DOCUMENT RESUME

ED 056 736 LI 003 281

AUTHOR	McAllister, Caryl
TITLE	A Study and Model of Machine-Like Indexing Behavior by Human Indexers.
INSTITUTION	International Business Machines Corp., Los Gatos, Calif. Advanced Systems Development Div.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
REPORT NO	Lab-R-16-205
PUB DATE	Nov 71
NOTE	146p.;(50 References)
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	*Indexes (Locaters); *Indexing; Information Retrieval; Information Scientists; Librarians; Methods; *Models; Scientists; *Subject Index Terms; *Technical Reports
IDENTIFIERS	*Machine Aided Indexing

ABSTRACT

Although a large part of a document retrieval system's resources are devoted to indexing, the question of how people do subject indexing has been the subject of much conjecture and only a little experimentation. This dissertation examines the relationships between a document being indexed and the index terms assigned to that document in an attempt to quantify the extent of "machine-like" indexing occurring when librarians and scientists index technical text. A number of possible relationships between the text and the index assignments are predicted and tested with two models: a multiple linear regression model and a Boolean combinatorial model. It is concluded that indexers in general do not index technical text in a "machine-like" fashion and that neither model is useful as a general predictor of human indexing. (Author/NH)

ED0 56736

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Caryl McAllister
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

16.205/November 1971
Laboratory Report

A STUDY AND MODEL OF MACHINE-LIKE INDEXING BEHAVIOR BY HUMAN INDEXERS

Caryl McAllister*

International Business Machines Corporation
Advanced Systems Development Dvision
Los Gatos, California

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ABSTRACT

This dissertation examines the relationships between a document being indexed and the index terms assigned to that document in an attempt to quantify the extent of "machine-like" indexing occurring when librarians and scientists index technical text. A number of possible relationships between the text and the index assignments are predicated and tested with two models: a multiple linear regression model and a Boolean combinatorial model. It is concluded that indexers in general do not index technical text in a "machine-like" fashion and that neither model is useful as a general predictor of human indexing.

This study was done in partial fulfillment of the requirements for a Ph.D. at the University of California at Berkeley. The work reported was supported partly by the IBM Work-Study Program and partly by a fellowship under Title II-B of the Higher Education Act of 1965. All computer time and office support were provided by the IBM ASDD Laboratory in Los Gatos, California.

LOCATOR TERMS FOR THE IBM SUBJECT INDEX:

Computer Simulation of Human Indexing
Information Science
Automatic Indexing
05 Computer Application

---

*Now with the IBM World Trade Laboratory in Böblingen, Germany

LI 003 281

Copyright © 1971

by

Caryl McAllister

TABLE OF CONTENTS

Acknowledgements v

Abstract vi

1. Introduction 1

2. A Review of Indexing Rules for Humans and Machines 7

    2.1 Rules for Indexers 8

    2.2 Results of Human Indexing 14

    2.3 Rules for Automatic Indexing 21

        2.3.1 Syntactic Clues 21

        2.3.2 Statistical Clues 23

        2.3.3 Textual Clues 24

    2.4 Assignment Rules 27

    2.5 Clues and Assignment Rules Used in this Thesis 29

        2.5.1 Regression Model Clue Types 29

        2.5.2 Combinatorial Model Clue Types 34

        3.5.3 Assignment Rules Used in the Models 36

    2.6 Subject Experts versus Librarians 39

3. The Multiple Linear Regression Model 43

    3.1 Desirable Characteristics of an Indexing Model 44

    3.2 The Mathematics of Regression 46

    3.3 The Correlation Coefficient 52

    3.4 Relative Importance of Clues 56

    3.5 Prediction with the Regression Model 57

    3.6 Computer Program for Multiple Linear Regression 59

4. The Combinatorial Model 61

    4.1 Reasons for the Combinatorial Model 62

    4.2 Types of Indexer-Model Agreement 63

    4.3 Boolean Combinations of Clues 68

4.4 Statistical Tests of the Combinatorial Model 73
4.5 Computer Programs for the Combinatorial Model 75
5. Experimental Procedures and Samples 79
5.1 The Documents and Indexers 80
5.2 Clue Counting Procedures 83
5.2.1 Keypunching 83
5.2.2 Reduction to Singular Form 84
5.2.3 Stemming 85
5.2.4 Thesaurus Terms Used in the Models 86
5.2.5 Document-Clue Matching Procedure 87
5.3 Sub-Samples Tested 88
5.4 Statistics Describing the Documents and Indexers 90
6. Conclusions 100
6.1 Introduction 101
6.2 Evidence For and Against Machine-Like Indexing 102
6.2.1 Results of the Multiple Linear Regression 102
6.2.2 Results of the Combinatorial Model 107
6.2.3 Comparison of the Results of the Models 109
6.3 Relative Importance of the Clue Types 112
6.4 Individual Documents and Indexers 117
6.5 Some Suggestions for Further Research 120
Appendix. Changes to Lovin's Stemming Procedures 128
Bibliography 132

Acknowledgements

A number of people have lent a helping hand during the last few years. I owe special thanks to some outstanding librarians: Jeanette Wood and Margaret Martin who first infected me with that disease called librarianship, and Marjorie Griffin who encouraged the virus.

My thesis committee: Dr. Patrick Wilson, Dr. William S. Cooper and Dr. C. West Churchman have been most helpful on a very tight schedule. I thank them for their time, valuable suggestions, and their co-operation.

Friends at IBM have supplied information and consulting services: Dr. A. Stratton McAllister (programming consultant, editor and official hand holder), Dr. Calvin K. Clauer (human factors and experimental design), Alan S. Neal (human factors and statistics) and Dr. Peter Bryant (statistics). The IBM ASDD Laboratory in Los Gatos, California has been very generous with computer time, office support and library service. I also owe much to Dr. Michael D. Cooper and to Theodora Hodges who served as general advisors during my years at Berkeley.

Librarian, scientist and engineer friends served as indexers. Their promptness and cooperation is much appreciated.

Needless to say, any errors are to be credited to my account.

Abstract

A Study and Model of Human
Indexing Behavior

Caryl McAllister

This dissertation examines the relationships between a document being indexed and the index terms assigned to that document in an attempt to quantify the extent of "machine-like" indexing occurring when librarians and scientists index technical text.

A number of possible relationships between the text and the index assignments are predicated and tested with two models: a multiple linear regression model and a Boolean combinatorial model. The models test two classes of relationships for the best relationship in that class. Both models find and correlate textual evidence in the document for a given index term with the descriptors assigned by the indexers. In all, some sixty types of textual evidence (or clues) are considered.

For the experiment twelve indexers were divided into two groups of six each: professional librarians and engineers or scientists. Each subject indexed all twenty sample documents. There was a significant difference between the amount of librarian indexing and the amount of engineer/scientist indexing accounted for. Although the difference was not great, the engineers and scientists proved

to be less predictable than the librarians on the basis of the textual clues.

Over the entire sample of documents and for all indexers, the regression model accounts for about 30% of the indexing. For a single document, however, as much as 40 to 80% of the indexing can be explained by the regression model. The location and type of textual clue deemed important by the indexers varies considerably from document to document. Hence variations in clue "style" among documents lowers the overall percentage because the entire sample is a compromise position for all the documents.

Regressions run on four single indexers show a very small correlation between clues and indexing ranging from 7 to 22%. Individually the indexers are less predictable then the group.

The information from the Boolean combinatorial model is less comprehensive primarily because not enough computer time was available for a full development of the model. Based on a one-third sample, the model correctly predicted about 65% of all indexing decisions. No other combinatorial runs were made.

It is concluded that indexers in general do not index technical text in a "machine-like" fashion and that neither model is useful as a general predictor of human indexing.

Chapter One

INTRODUCTION

1. Introduction

The information explosion is a widely recognized phenomenon. Increasing numbers of people engaged in research have produced increasing numbers of papers reporting that research. Libraries, engaged in the business of making that research available on demand, must process increasing numbers of such documents. This processing remains a major library bottleneck.

In addition to the investment in clerical labor and paperwork to acquire a document, a library often must also spend professional labor indexing it. This indexing makes it possible for patrons to find the particular items they want in a large collection without having to read the entire collection. The document and the index entries for a document are stored in some convenient place so that someone wishing to use the library or information center may search the indexes to locate it.

Two tools have been developed to aid indexers: indexing rules and lists of approved index headings. While both rules and headings are commonly available to aid in author indexing, subject indexing is quite another story. Here, lists of approved headings (also called thesauri) are plentiful, but there are only vague and imprecise notions of how an indexer should go about choosing the most appropriate headings out of the approved list for the document at hand.

Even though a large part of a document retrieval system's resources are devoted to this task, the question of how people do subject indexing has been the subject of much conjecture and only a little experimentation. There have long been arguments in the literature about the educational requirements for indexers. If indexers do little more than copy words from the document, we shouldn't be paying graduate-level subject experts to do the job. On the other hand, if indexers are involved in some rather sophisticated decision-making, we shouldn't be talking so glibly about substituting machines for people.

Only in biomedicine has anyone attempted even a partial answer to the question of how people go about indexing. Yet none of the biomedical studies has been conclusive enough to answer the question even for that particular field. And no-one has tried the experiment for less idiosyncratic literature than medicine.

For some time, researchers interested in automatic indexing have been proposing that machines should choose index terms on the basis of machine-recognizable textual clues present in the text. Such clues as noun phrases, word frequency or location, word stems and synonyms have been suggested. If textual clues account for a large part of a human indexer's behavior, then it might be feasible to automate indexing. And if this behavior can be modelled, the model could form the basis for just such an automatic indexing system. If, on the other hand, mechanicallyrecognizable clues do not account for a large part of a human indexer's behavior, automatic indexers would have to go beyond simple textual clues to do human-like indexing.

Because of the strong interest in machine-recognizable textual clues for automatic indexing, because of the numerous suggestions that human indexers do little more than word matching, and because a very large proportion of any reference retrieval system's budget is invested in indexing, this thesis attempts to answer the question: *To what extent do machine-recognizable textual clues account for human indexer behavior*?

To highlight the influence of training on indexing, we use indexers of two kinds: librarian-indexers, who by training and experience ought to know how to go about indexing, and scientist-indexers, who by training and experience ought to be most familiar with the subject matter to be indexed. Differences in indexing behavior between the two groups are of interest. We are also interested in the textual clues themselves and attempt to isolate those clues which contribute most to the explanation of human indexing. To do this effectively, a large number of clues and selection rules are covered systematically.

Chapter 2 reviews previous studies of human indexing and the indexing rules that have been suggested for automatic indexing. Besides surveying commonly quoted human rules, this chapter points out that rules used by humans are not, in fact, rules but general behavioral guidelines. The discussion of previous models of human indexing behavior points out the strengths and weaknesses of these studies. The analysis of rules used for automatic indexers shows the variety of rules discussed in the literature and suggests the types of textual clues which should be accounted for in the indexing models. The textual clues and the assignment rules used in the two models are discussed in this chapter.

The two models developed in the thesis are presented in Chapters 3 and 4. The first is a multiple linear regression model chosen for its statistical and predictive properties. The second is a combinatorial model which is used to test many of the clues summarized in the second chapter. Each model has advantages and disadvantages. Taken together, they complement each other. Both models quantify the extent to which machine-recognizable textual clues account for indexer behavior. Either can act in a predictive manner. Chapter 3 presents the regression model, statistical tests for regression and the computer program used for regression. Chapter 4 presents the same information for the combinatorial model.

Chapter 5 discusses the experimental procedures and gives descriptive information about the experimental samples. The computer programs written to obtain and analyze the data and to calculate results are presented in some detail in this section.

The conclusions of the thesis and suggestions for further research are in Chapter 6.

Chapter Two

A REVIEW OF INDEXING RULES
FOR HUMANS AND MACHINES

2. A Review of Indexing Rules for Humans and Machines

2.1 Rules for Indexers

For the purposes of this discussion we must distinguish between a Procedure and a Guideline. A Procedure is a set of exact and detailed rules which invariably lead the performer to the same outcome provided he is given the same input. Most computer programs are Procedures because, given the same input they operate on this input in exactly the same way each time to produce exactly the same output. The performer of the Procedure need not be a machine, however. Suppose I give you instructions for getting to my house from San Francisco. These instructions might include taking certain roads, turning in a specified direction at certain intersections, and so forth. _If_ you follow these instructions, _then_ you will arrive at my house. There is a guarantee that if the directions (the procedure) are followed, the result (arriving at my house) is assured. Of course, there is no guarantee that everyone arriving at my house has followed the same directions to get there.

In contrast, Guidelines have no guaranteed outcome. A Guideline is a set of warnings or cautions which are not detailed enough to invariably lead the performer to the same outcome even when given the same input. For instance, I might tell you to: head South; if speed is essential, take the freeway; watch for signs; use a map. These Guidelines tell you to watch for signs, but don't say which signs. They suggest that a map might be helpful, but don't tell exactly how a map is to be used or how it might be helpful. Guidelines for getting to my house won't guarantee arrival and they certainly won't guarantee that everyone using them will get to my house in the same way.

Let us draw the analogy to human indexing. Mention is made in the literature of "indexing rules". These rules are, in fact, Guidelines, not Procedures. They do not guarantee that anyone who follows them will arrive at the same index set. Proof of this may be found in indexing consensus studies (Hooper (1965), St. Laurent (1966)) where the same instructions, thesaurus vocabulary and documents almost invariably lead to different index sets when used by different indexers or even the same indexer at different times. We will review some of these indexing guidelines here because they are important for understanding how indexers go about their task.

Based on experience with chemical literature in an industrial company, Carol Penn (1962) outlines indexing as the search for answers to four groups of questions. Penn says the indexer first asks "What information is in this document, how is it organized, and into how many intellectual components is it subdivided?" No procedure is given for deciding what constitutes "information", but Penn suggests that the indexer read the most condensed document statement first (the title), and then work toward the most narrative (the abstract, then paragraph headings, and finally the full document). This suggestion is, in part, a procedure because it tells the indexer where and in what order to look. It does not, however, tell the indexer what to look for or when to stop looking.

The second question is: "How are the overall document and each of its component subdivisions related to or identified with the current and anticipated activities of the users?" This is an identification of the information from the point of view of the user as well as the author. The terminology of the author is put into relationship with the accepted terminology of the user group. But the terms "component subdivisions", and "current and anticipated activities of the users" are not defined, nor are any instructions given for finding out just what these current or anticipated activities might be. In order to estimate potential usefulness, indexers would have to estimate the likelihood that a project might be undertaken. But to expect indexers to predict the course of scientific investigation is to turn them into managers of scientific projects. This second rule, therefore, serves primarily as a warning to indexers that the needs of the users are an important factor in a reference retrieval system.

The next question is: "How new, how reusable or how original is the information in each component?" Penn argues that if the indexers cannot judge which information the users will consider new and interesting, then the indexing depth will be too great or too shallow. This rule requires that the indexer know the state of knowledge of present and future system users. This is an obviously impossible condition, yet it points to a common-sense notion that indexers shouldn't be indexing the obvious. The difficulty lies in deciding what is and what is not obvious.

To answer the last question: "How should information be described?" the indexer rephrases the mental picture of the document into descriptors from the thesaurus. If, indeed, the document is understood, then the indexer does have some idea of what the author is saying - he has a mental picture of the subject(s) of the document. But this does not assure that two indexers will have identical mental pictures nor does it assure that the interpretation of this mental picture into index terms will produce identical results.

In general, then, Penn's rules are cautions to warn the indexer that the subject content of a document, the activities and the subject expertise of the users, and the thesaurus vocabulary of the system are important and should be considered when indexing. But these cautions do not consitute a Procedure.

Other published indexing rules are similar to Penn's. Bernier (1965, 326) suggests the following: 1) choose to index those subjects which are novel, emphasized, or extensively reviewed, 2) index to the maximum specificity warranted by the author, 3) choose those terms most frequently used in the field, 4) provide guidance (cross-references) among headings and from synonyms, 5) check all index entries for accuracy, 6) use modifying phrases to make subject terms more specific and to provide better guidance. Again, these are cautions to the indexer about subjects that are novel, the maximum level of indexing specificity, etc. Bernier's rules substantiate the fact that so-called indexing rules are not procedural. Rees (1962) and MacMillan and Welt (1961) agree.

We have pointed out the vagueness and imprecision inherent in the indexing "rules" to be found in the literature. The business of indexing is no more procedural when seen from a philosophical point of view. Wilson (1968) discusses several ways one might determine the subject of a document. For instance, an indexer might list, sentence by sentence, what a document was about. The list could justifiably include the names of the objects mentioned in each sentence, or the names of the concepts employed by the author in expounding on his subject, or the names of the things or individuals indirectly referred to, or any combination of these. While it is possible to recognize obviously wrong entries on this list, knowing what is obviously wrong does not resolve the many occasions when indexers can differ considerably in acceptable indexing assignments. Wilson's arguments point out, once again, that indexers are operating with Guidelines.

In conclusion then, we have seen that the indexing rules humans profess to use are not Procedures, but Guidelines.

Indexing rules may give general guidance; they do not constitute a how-to-do-it course. According to the dictates of the indexing profession, indexing is an art, not a science. Consciously, at least, human indexing involves a great deal of judgement, subject expertise, knowledge of the users and of the document retrieval system. None of these things is easily automated by present day standards of artificial intelligence.

This is not to say that we cannot use a procedure to mimic human indexing. As the next section demonstrates, what indexers do and what they say they do may be quite different things.

2.2 Results of Human Indexing

Instead of investigating what indexers say they do, some experimenters have tried to find out what indexers do by looking at the index sets produced. Studies of this kind cannot claim to have investigated the paths indexers used to arrive at a particular index set. However, possible hypothetical mechanisms for reaching a particular index set can be investigated and the outcome of these artificial rules can be compared with the outcome of human indexing.

Fels and Jacobs (1963) were interested in the extent to which indexers became "linguistically creative" when indexing. They defined three sources of indexing terms: 1) words occurring in the text, 2) synonyms for text words, and 3) paraphrases of the text. These types of index terms are increasingly "creative". Using random samples taken from state and federal statutes, straight term selections constituted 63 to 91% of the index set, synonym substitutions ranged from .5 to 5.8% and paraphrases from 7.4 to 33.7%. The statistics quoted indicate that legal indexers, at any rate, are not particularly creative linguistically. Note that although this study indicates where the indexing words came from, it does not indicate how the indexers arrived at particular index entries.

A study by Montgomery and Swanson (1962) strongly substantiates Fels and Jacobs. They chose subject headings at random from Index Medicus. Each of the titles indexed under each of these headings was compared with the heading itself. In 86% of the cases the subject heading or a synonym for it, appeared in the title.

O'Connor (1964) later disagreed with the Montgomery-Swanson study. He argued that it ignored subdivisions of subject headings and used synonyms inconsistently to obtain the high degree of matching. To substantiate his points, he tried the Montgomery-Swanson indexing rules on titles from three medical indexing systems. Based on samples of 50 titles from each of the systems, the heading-title correlation in these samples ranged from 19 to 45% in the first system, from 40 to 68% in the second, and from 13 to 39% in the third. This is in sharp contrast to the 86% agreement obtained by Montgomery and Swanson. At least as far as medical text is concerned, there is little agreement on the profitability of using title words and their synonyms as an artificial procedure for imitating human indexing.

A few studies have been made of indexing in engineering. Slamecka and Zunde (1963) found 80% of the humanly-assigned index terms in the abstracts of 30 documents from Scientific and Technical Aerospace Reports. Bottle (1970) compared the titles of articles with humanly-assigned subject headings for each article. Titles were chosen from Applied Science and Technology, British Technology Index and Engineering Index. From 48 to 68% of the titles either matched the assigned heading or contained a syntactic variant or a synonym for it. Graves and Helander (1970) compared titles and abstracts taken from Petroleum Abstracts with the humanly-assigned index terms. Exact and synonym matches accounted for 40% of the humanly-assigned index terms. Although each of these studies was in the same general subject area of engineering, the percentage of human index terms accounted for ranged from 40 to 80%.

The studies discussed up to this point investigated possible mechanisms for arriving at the same indexing humans produced. All of the studies worked from the already-assigned index set backwards to the text. In effect, this approach covers only half of the problem. It accounts for where the index term came from; it provides a textual justification for the assignment of each index term. But it does not tell how many matches with other subject headings might have occurred. For instance, suppose the title "Real-time Input Preprocessor for a Pattern Recognition Computer" were compared with the subject heading "Pattern recognition". There is an exact match between the subject heading and a portion of the title. This would be counted as one instance of an exact thesaurus-title match in the studies discussed above. But this same title also matches two other subject headings: "Real-time computer systems" and "Input preprocessors for computers". These matches were ignored by the above studies. Although these studies kept track of the index terms or subject headings which were assigned, they did not try to explain why other terms were not assigned. Both explanations are required in a complete model.

In later experiments, O'Connor tried several methods for obtaining manually assigned index terms from full document text (1961, 1962, 1965). He chose two index terms, 'toxicity' and 'penicillin', from the thesaurus of an operational 10,000-document system. He then tried to formulate rules for assigning the documents to the appropriate subject heading without assigning other documents in the collection to that subject heading. In the end, a quite complicated indexing rule was formulated for each thesaurus term. These rules, while assigning 'toxicity' to most of the toxicity papers, also assigned 'toxicity' to non-toxicity papers. To counteract the over-assignment, without causing concomitant under-assignment, O'Connor used minimum frequency requirements, location of the toxicity clue in specific parts of the document, etc.

The rules formulated on an initial group of toxicity papers were then tested on a second group of papers from the same system. They correctly selected 92% of the toxicity papers, at the cost of over-assigning 'toxicity' to 18% of the non-'toxicity' papers. The computer-simulated rules were comparable with the system's regular human indexers who correctly assigned about 80% of the toxicity papers with a 2% over-assignment.

A similar procedure was followed for the term 'penicillin' resulting in another set of simulated computer-assignment rules. These rules correctly selected 97% of the penicillin papers at the cost of a 4% over-assignment. In contrast, indexers correctly assigned about 75% of the papers and over-assigned less than 2%.

Although the artificial indexing rules O'Connor devised work quite well for 'penicillin' and 'toxicity', there are some difficulties with his scheme. First, because each thesaurus term requires a different rule, the invention, programming and use of such rules for a real-life thesaurus (say, 20,000 terms) is almost a practical impossibility. Second, the two sample index terms selected for study were both single words, and were posted on a rather high proportion of the collection's documents (1500/10,000 for toxicity and 700/10,000 for penicillin). Such heavy posting is most unusual and occurs on fewer than 2 or 3 percent of the terms even in very large collections (Houston and Wall (1964)). Third, the study was done on biomedical literature which typically has a well defined and very specific vocabulary. There is nothing comparable to O'Connor's list of disorders in the vocabulary of engineering, and one might expect indexing rules to be different when the vocabulary is less precise.

In conclusion, there are several major objections to most of the studies we have discussed: the particular human indexing chosen as a standard, the question of over-assignment, and the investigation of only a few possible artificial rules. In each of these studies, the human indexing which acted as the standard was not all done by the same person or group of people. This is an important point because of the effect it has on the rules the experimenter devises to account for the indexer's behavior. Let us suppose that two indexers have rather different indexing practices. One of them (Indexer One) assigns an index term only if an exact match for a thesaurus phrase occurs one or more times in the document. The other (Indexer Two) assigns the term only if the exact match occurs two or more times. Now suppose that Indexer One indexes X percent of the sample documents, and that Indexer Two does the remaining Y percent. The experimenter could come up with a rule which says "assign the term if it occurs at least two times in the document". This rule will omit up to X percent of the assignments. If the experimenter decides the rule should be "assign the term if it occurs one or more times in the document", then he will be over-assigning in up to X percent of the cases. If many indexers and many indexer assignment rules are involved, the hypothetical assignment rule devised by the experimenter is very dependent upon the particular mix of people who did the indexing.

There are two ways to deal with this problem. First, all the documents could be indexed by the same person. The experimenter would then be looking for a rule to explain the behavior of a single indexer. The second possibility is to have all the documents indexed by each of a group of people. This leads the experimenter to an explanation of an "average" type of indexing. Since an individual indexer is unlikely to be following any rule consistently, the averaging would give an opportunity for individual variations to cancel out.

The second major objection to most of the studies discussed above is that they have ignored or played down the effects of over-assignment. The artificial rule must account for the non-assignment of terms as well as the assignment of terms. This difficulty was discussed above in connection with the Fels and Jacobs and the Montgomery and Swanson studies.

Third, there has been no systematic investigation of a broad spectrum of possible hypothetical indexing rules. As Section 2.4 demonstrates, a combination rule (clue one AND clue two) is very seldom employed. Investigation of a broader range of rules would make it possible to say just how complex an artificial rule must be to imitate human indexing.

Despite the assurances of an occasional devotee (Salton (1970)), there is no clear evidence that human indexing is "machine-like". The models proposed in this thesis are intended to investigate two general types of machine-like rules to determine whether they do account for a large percentage of human indexing behavior.

2.3 Rules for Automatic Indexing

Hypothetical indexing rules have been suggested for purposes other than imitating human indexing; much of the literature on automatic or mechanical indexing consists of tests of such hypothetical rules. Instead of surveying the literature of automatic indexing which has been reviewed exhaustively and competently by Stevens (1970), we will try to summarize the types of rules proposed for automatic indexers.

The automatic indexing rules mentioned in the literature break down naturally into four general areas: 1) syntactic clues, 2) statistical clues, 3) textual clues and 4) assignment rules. In this section we will characterize the three types of clues and cite examples of each type. Section 2.4 discusses the assignment rules. We are primarily concerned with the textual clues and the assignment rules because they provide a basis for understanding the models used in Chapters 3 and 4.

2.3.1 Syntactic Clues

Syntactic analysis makes a first step toward understanding the meaning of text by unravelling the text's grammatical structure. Syntactic clues are chosen on the basis of knowledge of this grammatical structure. An automatic analyzer finds the part of speech of each word in the text as it parses the sentence. Unfortunately, this is not a simple process. Syntactic analyzers are often quite complicated programs which can produce a number of alternate readings of a single sentence. Dealing with two sentences is beyond the abilities of most existing programs unless the vocabulary and grammatical structures are severely limited. Although Harris (1959) talked of kernalization of sentences and replacement of pronouns in 1959, only recently have there been programs which can actually perform some of these feats (Shapiro, et.al. (1969)). In fact, artificial intelligence experimenters count the understanding of small portions of text about calculus a major success (Simmons (1970) 21) mainly because of syntactic problems.

There have been automatic indexing experiments with syntactic analyzers designed to search for specific types of syntactic clues, however. Baxendale (1958, 1962), Baxendale and Clarke (1966) and Clarke and Wall (1965) identified noun phrases in natural language text with an accuracy of 91%. Unfortunately, this program has never become part of an automatic indexer. Klingbiel (1969, 1971) designed a program to read in natural language text, locate phrases which could serve as potential index terms, and display these phrases to a human indexer. The human was expected to make the final indexing decision. This analyzer recognized just thirteen syntactic types.

While syntactic information will no doubt be an important automatic indexing technique in future years, for the present it is more talked-about than practiced. This clue type is not included in either of the models in this thesis.

2.3.2 Statistical Clues

The statistical methods of isolating clues are really methods for locating content-bearing words in natural language text. Large quantities of text must be processed - usually by truncation and counting - to give statistical information about the frequency of occurrence of text words in the language as a whole. The object is to locate words which have atypical distributions in the text.

For instance, Dennis (1965, 1967), in one of the earliest statistical experiments dealing with text, tested a number of statistical distributions intended to separate content-bearing words from the other words. About 3.8 million words from 2600 reports of law cases were keypunched. Then a number of statistical distributions were tested against this text to find one which characterized the content-bearing words. The content-bearing words identified by the distribution became the master indexing list. Every time one of these words appeared in a document, the document was assigned that word as an index term.

Damerau (1965) performed similar experiments with one million words of world politics news broadcasts. The object was the same: to find a statistical distribution which would accurately separate content-bearing words. He found that non-content-bearing words (often called "function" words) had a Poisson distribution through the documents since these words tended to be randomly distributed. Later, Stone (1967) and Stone and Rubinoff (1968) tried several modified Poisson distributions on a 70,000-word sample taken from Computing Reviews. Stone found that words with a Poisson distribution, since they occur randomly, are non-specialty or uninformative words. Specialty words have non-random, non-Poisson distributions. Stone developed two Poisson formulas and proved that one of them is analogous to Dennis' best separating formula.

The identification of content-bearing words is a first step in the compilation of a list of keywords. And a list of keywords can be very useful when building a thesaurus. But such a list does not, in itself, act as an automatic indexer. For this reason, statistical methods of isolating clue words are not included in either thesis model.

2.3.3 Textual Clues

Textual clues (also called 'machine-recognizable textual clues' or, simply 'clues' in this thesis) are the most common raw material for automatic indexing algorithms. Textual clues are words or phrases produced by natural language text or obtained from it *without* benefit of syntactic analysis or statistical manipulations of large quantities of text. Since this is a definition-by-default, some examples might be helpful.

Many years ago Luhn (1957) suggested the use of location as a textual clue. Words occurring in the title were supposed to be more likely to be good descriptors than words occurring in the body of the document. Other suggestions have been made for locations of textual clues. Baxendale (1958) thought the first and last sentences in each paragraph were good. O'Connor (1965) tried the first and last paragraphs of a document. Figure 2.01 lists the various locations or combinations of locations tried by various experimenters and references the journal article in which each suggestion was made.

A second group of textual clues centers around a match between the text of the document and a word list of some sort. By far the most common type of match sought is an exact match between the document and a word list or thesaurus (see Figure 2.01). Fangmeyer and Lustig (1969) and Montgomery and Swanson (1962) accepted a partial match between the document and the word list. Other experimenters searched for stems of words, or utilized thesaurus cross-references as clues.

The last major group of textual clues is based on counting. Here, a count of the number of times a word is used in a document determines whether that word is a clue or not. Some experimenters (see Figure 2.01 again) simply take the most frequently used words. Others take words occurring at least X times in a document, or those which constitute at least X% of the document. This counting procedure is to be contrasted with the procedures used to obtain statistical clues. Statistical clues are only available from large quantities of text (on the order of a million words). The counting procedure discussed above operates only on the document at hand. It does not depend on statistical word distributions in the language as a whole.

Each of the methods in these three major groups of textual clues is a way to obtain information about the subject content of the document from its text. The two other methods discussed in Sections 2.3.1 and 2.3.2 for obtaining information about subject content (syntactic clues and statistical clues) require either rather complicated programming or large quantities of text. The textual clues mentioned here are by far the most numerous clue types found in automatic indexing experiments - probably because they are the easiest clues to obtain with present-day computers. For this reason, they are the clues modelled in this thesis.

2.4 Assignment Rules

An automatic indexing algorithm is a combination of two elements: the clues identified, and the assignment rules. Given a particular pattern of clues, the assignment rule decides whether those clues result in an index term. For example, suppose the clue-finding procedure looks for thesaurus words in the document in two places: the abstract and the title. An assignment rule might be the following: "Assign the index term if the thesaurus word occurs once in the title or at least three times in the document". The assignment rule keeps track of the locations, frequencies and types of clues appearing in the document. When the minimum assignment rule conditions for a particular thesaurus term are met, that term is added to the document's index set. The assignment rule is simply an indexing procedure operating on textual information about the documents.

Many studies have made use of very primitive assignment rules. The most common of these is: if any textual clue occurs, then assign the corresponding index term (see Figure 2.02). In some cases, several textual clue types are involved. For instance, Artandi (1969) looked for two significant words in the same sentence. Montgomery and Swanson (1962) searched for at least one of several clue types. Luhn (1957) searched for words in particular locations with high frequencies. O'Connor (1965) developed increasingly more complicated assignment rules for two index terms in the medical field. In fact, his assignment rules were different for each index term studied.

In conclusion, we have seen that a number of hypothetical indexing rules have been proposed and tested in the pursuit of automatic indexing algorithms. Unfortunately for us, the results of these automatic indexing experiments are sometimes not evaluated at all, are evaluated only in terms of the total number of terms in the index set, or are compared with the output of a single human indexer. Although none of the experimental results are particularly useful to us in deciding what proportion of human indexing can be accounted for by textual clues, these studies do give us valuable insight into hypothetical rules which could be used to imitate human indexing.

2.5 Clues and Assignment Rules Used in this Thesis

The clues and assignment rules modelled in this thesis are extensions of those found in the literature (see Figures 2.01 and 2.02) with adaptations to accommodate the documents actually used. For instance, since the sample documents indexed are short and consist of just a title and abstract, just two locations for the clues are distinguished: title and abstract. On the other hand, extensive use is made of information from the thesaurus for identifying clues. Sections 2.5.1 and 2.5.2 describe and define the clues for the regression and combinatorial models. Section 2.5.3 describes the assignment rules typified by the two models.

2.5.1 Regression Model Clue Types

In keeping with the breakdown found in the literature, clues have been divided into three general groupings:

1 type of match (6 different types in group)

2 length of match (5 different lengths in group)

3 location of match (2 different locations in group).

One element is taken from each of the three groupings to constitute a single clue. For example, a main entry descriptor match (group one) of a three-word phrase (group two) in the abstract (group three) is a single clue. There are 6•5•2 or 60 possible clue types.

The three short lists below constitute a complete display of each of the items in the groups. All possible clues are formed by taking every possible combination of matches from the three groups.

Type	Length	Location	
main entry	three-word phrase	title	
stem	two-word phrase	abstract	
used-for term	header		(2.01)
broader term	modifier 2		
narrower term	modifier 1		
related term			

These sixty clue types may be thought of as a sixty-place string of numbers. The position of the number in the string indicates the clue type, the value of the number itself is the frequency of occurrence of that clue type. For example, the first number in the string of numbers is the position for three-word main entry descriptor matches in the title. If a '2' occurs in this location for a given document, there are two three-word main entry phrase matches for the thesaurus term in the title of the document. We call this sixty-place string of numbers a "clue vector". There is a clue vector for each document-term pair analyzed. These clue vectors form the basis of the multiple regression model discussed in Chapter 3.

Each of the matches is operationally defined by the computer programs used to isolate it. A definition of what constitutes a match between the document and the thesaurus phrase is given below. Information on the computer programs may be found in Chapter 5.

To understand what is meant by each component of a clue type, consider the following excerpt from a thesaurus.

Radiation counters
BT Measuring instruments
Radiation measuring instruments
NT Beta spectrometers
RT Dosimeters
Ionization chambers
Vertical takeoff aircraft
UF convertiplanes

where BT = broader term, NT = narrower term, RT = related term, and UF = used for.

Main entry: the thesaurus and the document word(s) match exactly, character for character. A singular/plural difference is counted as an exact match. Thus 'counters' in the thesaurus matches 'counter' or 'counters' exactly.

Stem match: the stem of the thesaurus word and the stem of the document word(s) match exactly. The stem of a word is that part of a word to which inflectional endings are added or in which phonetic changes are made for inflection. The thesaurus stem 'radia' matches the document stem 'radia' for such unstemmed words as 'radiation', 'radiate', etc.

Used-For match: the UF references in the thesaurus match either the singular or the plural form of the word(s) in the document. A used-for match is counted for the thesaurus term 'vertical takeoff aircraft' if either 'convertiplanes' or 'convertiplane' occurs in the document.

Broader term match: the BT references in the thesaurus match either the singular or the plural form of the word(s) in the document. A broader term match is counted for the thesaurus term 'radiation counters' if 'measuring instruments' or 'measuring instrument' or 'radiation measuring instruments' or 'radiation measuring instrument' occurs in the document.

Narrower term match: the NT references in the thesaurus match either the singular or the plural form of the word(s) in the document.

Related term match: the RT references in the thesaurus match either the singular of the plural form of the word(s) in the document.

Three-word phrase: if the thesaurus term being tested is a three-word phrase, and the words occur in the document with no more than one intermediate 'of' then a three-word phrase match has occurred. A three-word phrase match for 'vertical takeoff aircraft' occurs if either 'vertical takeoff aircraft' or 'takeoff of vertical aircraft' or 'vertical takeoff of aircraft' or 'aircraft vertical of takeoff', etc. occur in the document.

Two-word phrase match: if the thesaurus term being tested is a two word phrase, and the words occur in the document with no more than one intermediate 'of' then two word phrase match has occurred.

Header match: if the right-most word of a multi-word thesaurus phrase occurs in the document, or if a thesaurus entry of a single word occurs in the document, then a header match is counted. If either 'dosimeters' (a thesaurus entry of a single word) or 'chambers' (the right-most word of 'Ionization chambers') occurs in the document, a header match is counted. If a thesaurus term of the form 'card punches (data processing)' occurs, the parenthesized expression is ignored. In this case 'punches' is the right-most word of a two-word phrase and is therefore the header.

Modifier 2 match: if the second word of a three-word thesaurus phrase, or the left-most word of a two-word thesaurus phrase occurs in the document, then a modifier 2 match is counted. A modifier match for 'vertical takeoff aircraft' is counted if 'takeoff' occurs in the document; a modifier 2 match for 'radiation counters' is counted if 'radiation' occurs in the document.

Modifier 1 match: if the first word of a three-word thesaurus phrase occurs in the document, then a modifier 1 match is counted. The word 'vertical' is a modifier 1 match for 'vertical takeoff aircraft'.

Title match: if the word(s) being matched occur in the title, then a title match has occurred.

Abstract match: if the word(s) being matched occur in the abstract, then an abstract match has occurred.

The textual clues occurring in the document may be counted more than once. If both 'vertical takeoff aircraft' and 'aircraft' occur in the abstract, this counts as one exact three-word phrase match in the abstract and two exact header matches in the abstract. This method of counting assures that each clue is counted independently of all others.

2.5.2 Combinatorial Model Clue Types

The regression model in Chapter 3 and the combinatorial model in Chapter 4 have been tested with the same clue types. However, the additive properties of the regression model and the Boolean properties of the combinatorial model require somewhat different reporting schemes for these clues. The regression model simply records the count of the number of times a clue appears in the document. The combinatorial model uses Boolean combinations, so the numbers in the clue vector must be binary (either one or zero). This is accomplished by translating the single-cell count of the regression model into a binary record. There is a zero in the binary record if there is a zero in the corresponding place in the regression model record. There is a one in the binary record if there is a number greater than zero in the corresponding position of the regression model. The binary vector simply records which clue types are present in the document. A zero value in a binary clue cell means the clue type did not occur in the document; a one means that one or more clues of that type occurred in the document.

This particular pattern for the binary clue vector was chosen for two practical reasons: 1) for the size of documents used in the sample, there is little necessity to record broad frequency ranges since high frequency clues are not common and 2) additional clue types increase computational time considerably. In theory there is no limit to the occurrence frequencies which could be represented by a binary record, however. As with the regression model, there is one clue vector for each document-term pair analyzed.

The following three short lists summarize the clues used for the combinatorial model.

Type	Length	Location	
main entry	three-word phrase	title	
stem	two-word phrase	abstract	
used-for term	header		(2.02)
broader term	modifier 2		
narrower term	modifier 1		
related term			

These lists are identical to those in Equation 2.01 except for the the modification of the options in the location group.

As with the regression model, all possible clues are formed by taking every possible combination of matches from the three groups. There are a total of 6•5•2 or 60 possible clues.

2.5.3 Assignment Rules Used in the Models

The models in Chapters 3 and 4 are intended to test a number of possible assignment rules in a systematic fashion. Each model tests a different class of assignment rules although in a certain number of special cases the two kinds of assignment rules are mathematically equivalent.

The class of assignment rules tested by the combinatorial model are a particular set of Boolean equations formed from combinations of the sixty binary clues. These Boolean equations are of the form (clue-type-1 AND clue-type-2) OR (clue-type-3) OR (clue-type-4 AND clue-type-5). Translated into a model of human indexing, the above equation would read: if clue-type-1 AND clue-type-2 OR if clue-type-3 OR if clue-type-4 AND clue-type-5 are present in the document, then assign the thesaurus term. These equations are covered in more detail in Chapter 4.

The class of assignment rules tested by the multiple linear regression model is of a different form: $Y = A + B_1$ (number of clue-type-1-occurrences) + B_2 (number of clue-type-2 occurrences) + ... + B_n (number of clue-type-n occurrences). Translated into a model of human indexing, this equation would read: To a constant, A, add the coefficient B_1 multiplied by the number of times clue-type-1 occurred; then add the coefficient B_2 multiplied by the number of times clue-type-2 occurred; etc. The sum Y is the percentage of indexers assigning the term. The multiple linear regression model looks for additive combinations of the textual clues. Each clue is weighted arithmetically by the coefficients so the total score for a particular term is a sum of the fractions of all of the clues considered. (See Section 3.2 for a detailed discussion of this weighting.) The object of the regression calculations is to find the "best" values for the constant and coefficients. Chapter 3 discusses the regression in more detail.

As mentioned above, in a certain number of special cases, the Boolean combinatorial model and the multiple linear regression model are equivalent. A branch of switching theory, called "threshold logic", deals with this equivalency. Threshold logic (Lewis and Coates (1967)) is concerned with converting binary circuits (or Boolean equations) into threshold circuits (or a sequence of linear equations). A number of methods are available for "realizing" (converting from Boolean to) a threshold logic element. All Boolean equations can be realized by one or more threshold logic elements. However, only a few Boolean equations may be converted to a single linear equation. When just a single threshold element is needed, the Boolean equation is said to be "linearily separable". If there are two Boolean variables (in our case a Boolean variable is a clue type), then there are 16 distinct Boolean functions of which 14 are linearily separable. If there are three Boolean variables, then there are 256 distinct functions of which 104 are linearily separable. When the number of Boolean variables is equal to or greater than 4, the percentage of linearily separable functions decreases rapidly (Torng (1966) 20). The equivalency between the best Boolean and the best regression models is discussed in Section 6.2.3.

Since the regression model does not permit testing of many Boolean selection rules because of the low density of linearily separable functions, a Boolean combinatorial model is also desirable. In this thesis, one particular group of Boolean assignment rules is tested exhaustively to uncover the best set of Boolean equations for the sample documents.

Both models assume that the same indexing procedure or assignment rule applies to all terms in the thesaurus. This is consistent with the approach taken by all automatic indexing studies with the exception of O'Connor who devised a different rule for each thesaurus term. Both models are, of course, dependent upon the particular clue types chosen by the experimenter. Neither model can disclose the importance of clue types not included in the model.

2.6 Subject Experts versus Librarians

If indexers do little more than pick good words out of the document, then a high level of subject competence may not be necessary. On the other hand, if indexers make intellectual decisions requiring knowledge about technical subjects, potential users of the system, etc., then subject expertise is an obvious prerequisite.

Although comparative studies have been made of author-indexers versus professional indexers, no comparison has been made of the dependence of the two groups on the textual clues in the document. One would expect that scientist-indexers would depend less on the actual words used in the documents because of their greater understanding of the subject matter. Librarian-indexers would not have the benefit of subject familiarity and would, therefore, be more dependent upon the words actually used in the document when indexing.

To test this hypothesis, two groups of indexers have been used as subjects for this study. The first group consisted of six librarian-indexers. Each of the librarians had an M.L.S. degree from an accredited school. Each had spent some time either indexing or cataloging in a special library in the field of engineering or science. Each had worked on a reference desk answering questions from patrons of the same kind of library. Each was familiar with the standard scientific and engineering abstracting journals.

The second group consisted of six scientist-indexers. Each of these scientists or engineers had at least an undergraduate degree in engineering or the hard sciences. In some cases, the scientist had an M.S. or a PhD. Each was earning a living as a scientist or engineer at the time of the study. The documents used for the experiment were in the field of instrumentation. This topic was chosen because scientists and engineers familiar with that subject were available to do the indexing.

Figure 2.01 Table of Textual Clues

A. Researchers using location as a clue:

- title, abstract, headings, text, references, figures
    - Edmundson and Wyllys (1961)
- first and last paragraphs
    - Luhn (1957) 315
    - O'Connor (1965) 499
- title and first paragraph
    - Swanson (1963)
- title, abstract, full text
    - Luhn (1959)
- first and last sentences in paragraph
    - Baxendale (1958)

B. Researchers using type of match as a clue:

- thesaurus or word list matches
    - Artandi (1964, 1969)
    - Bloomfield (1966)
    - Fangenmeyer and Lustig (1969)
    - Harris (1959)
    - Luhn (1959)
    - Meyer-Uhlenried and Lustig (1963)
    - Montgomery and Swanson (1962)
    - O'Connor (1965)
    - Salton (1968) 26
    - Slamecka and Zunde (1963)
    - Swanson (1960)
    - Zunde (1965)
- part of a thesaurus phrase
    - Fangenmeyer and Lustig (1969)
    - Montgomery and Swanson (1962)
- cross-references from the thesaurus
    - Fangenmeyer and Lustig (1969)
- stem matches
    - Fangenmeyer and Lustig (1969)
    - Luhn (1958)
    - Salton (1968) 30-33
    - Zunde (1965)
- multi-part clue expression with variable substitutions
    - O'Connor (1965)

C. Researchers using count and frequency criteria as clues:

- absolute frequency counts
    - Baxendale (1958)
    - Jones, Giuliano and Curtice (1970)
    - Luhn (1958)
- relative frequency counts
    - Artandi (1969) 218
    - O'Connor (1965) 499, 508
- most frequent words
    - Luhn (1957, 1958)
- most frequent word pairs
    - Baxendale (1958)
    - Edmundson and Wyllys (1961)

Figure 2.02 Table of Assignment Rules

a match with the thesaurus or with a word list
    Artandi (1969)
    Bloomfield (1966)
    Fangmeyer and Lustig (1969)
    Harris (1959)
    Jones, Giuliano, Curtice (1970)
    Montgomery and Swanson (1962)
    Salton (1968) 25-48
    Zunde (1965)

most frequent words in first and last sentence
    of each paragraph
    Baxendale (1958)

no more than X non-significant words separating
    significant words
    O'Connor (1965)

two significant words in the same sentence
    Artandi (1969) 219

two significant words within two paragraphs
    Luhn (1957)

at least X occurrences of thesaurus words per
    Y words of text
    O'Connor (1965)

title, heading, resume and frequency
    Luhn (1957)

Chapter Three

THE MULTIPLE LINEAR REGRESSION MODEL

3. The Multiple Linear Regression Model

3.1 Desirable Characteristics of an Indexing Model

An ideal model of textually-clued indexing would have several properties. First, it should answer the question "How strong is the relationship between the clues in a document and the index terms assigned to that document?" The answer to this question would tell us just how much of the indexing can be accounted for on the basis of the clues.

Secondly, it should be possible to make some statistically valid statements about the entire population of indexers and documents with the information obtained from the single sample. We would like to be able to infer that the relationship found in the sample also holds for the population as a whole.

Thirdly, the model should be able to be used predictively. It should say whether a particular index term would be assigned to an arbitrarily chosen document. This prediction might not be just a yes/no decision, but could also be, say, a prediction of the percentage of indexers who would assign the term to the document. If it turned out that there were only a small statistical relationship between the clues and the indexing assignments, then this predictive property would not be of much practical importance since the model could not function in place of the real indexers. If, however, there were a strong statistical relationship, a predictive model could be substituted for the indexers.

Because of the capability of giving strong answers to these requirements, multiple linear regression has been chosen as our first indexing model. Since this model assumes a linear relationship between the index terms assigned and the clues, a second model has also been built. This model, called the combinatorial model, does not assume linearity. The multiple linear regression model will be discussed in this chapter and the combinatorial model in the next.

3.2 The Mathematics of Regression

This section gives a cursory explanation of multiple linear regression. Although many statistics texts treat the subject, most discussions are difficult to read. The following books may be consulted for more detailed discussions: Hays ((1963) 490-577), Ferber ((1949) 346-379), Ostle ((1963) 159-243), and Draper and Smith (1966).

Regression is a common statistical technique used to show the linear relationships among two or more variables. For instance, we would like to know whether the index terms assigned to a document are related to the occurrence of textual clues in the document. In this case, the dependent variable is the percentage of indexers who assign a given index term and the independent variables are the various types of machine-recognizable textual clues in the document.

Assume for the moment that several indexers individually choose index terms from a thesaurus for the same document. In effect, the indexers are voting for the set of most popular index terms from among the potential thesaurus candidate terms. Some index terms will receive many votes, others fewer, most will receive no votes at all. Each of the potential thesaurus candidate terms considered by the indexer group is a single experimental event. This experimental event consists of the (n+1) numbers:

1 number of times clue type 1 occurred in document,

2 number of times clue type 2 occurred in document,

.

.

.

n number of times clue type n occurred in document,

n+1 percentage of indexer group voting for term.

For example, let us suppose the document indexed has the word 'computers' in it twice and that the index term now being considered is 'computers'. If clue type 7 is the exact match between the index term and a word in the document, then clue type 7 occurs twice in this document; therefore the number in the seventh place in the (n+1)-tuple is a 2. The numbers 1 through n form the clue vector discussed in Section 2.5.1. The clue types used in the model are also listed in that section.

Each experimental event is represented numerically by an (n+1)-tuple where n is the number of known document characteristics. In this case, n is the number of types of machine-recognizable textual clues tested by the experiment. The remaining point in the (n+1)-tuple is the dependent variable or the percentage of indexers assigning that term to the document.

As each of the potential thesaurus candidate terms is considered in turn, a new (n+1)-tuple is produced to represent the differing percentages of indexers who assign the term and the different quantity of textual clues in the document for that term. If all indexers index the same documents with the same thesaurus, then there will be

$$N = (\text{documents indexed}) \cdot (\text{size of thesaurus}) \qquad (3.01)$$

experimental events or (n+1)-tuples.

Each of these experimental observations can be represented in (n+1)-space as a single point. The object of the multiple linear regression is to fit the best straight line through these points. This line is fitted so that the summed squared deviations of the points from the line are minimized.

The equation of the resulting straight line is the classic one:

$$Y = A + B_1X_1 + B_2X_2 + \ldots + B_nX_n \qquad (3.02)$$

where A is a constant, the X's are the n clue types, and Y is the proportion of indexers assigning the term.

The B's can be thought of as weights for each clue type in the regression equation. Equation 3.02 can be re-written as:

$$
\begin{aligned}
Y = A &+ B_1 \bullet (\text{clue type } 1) \\
&+ B_2 \bullet (\text{clue type } 2) \\
&+ \ldots \\
&+ B_n \bullet (\text{clue type } n).
\end{aligned}
$$

Here the B's weight the clue types so that the sum of each of the terms in the equation totals to Y.

Notice the additive nature of the effects of the various clue types. This model says that an indexing decision is based on a weighted sum of all clue types, each clue type adding its evidence to the total evidence available for that index term. This assumption of linearity is basic to the regression model. It allows us to find the single best-fitting straight line for the data.

The use of multiple linear regression requires two assumptions about the data. These assumptions are not needed to calculate the correlation coefficient, but are required to say how good the correlation coefficient is as an estimator of the true population coefficient and to set confidence intervals. The first of these assumptions, normality, says that for a given X value, the Y values are distributed normally about a mean. When only a single X value is involved, or the values of X's can be controlled by the experimenter, the data can be inspected to see whether the assumption of normality is justified. Since our model has many X's whose values are not under experimental control, it is very difficult to determine how the Y values are distributed. It turns out, however, that deviations from normality do not have a serious influence on the regression model (Scheffe (1959) 350,360-368). Regression is not very sensitive to non-normality.

However, the upper and lower bounds set on the correlation coefficient are very dependent upon homoscedasticity. The homoscedasticity of a variable is the degree to which its variance is constant; that is, the degree to which the variance of Y given X is the same for all X. Unequal variances play havoc with the setting of confidence intervals. One way to deal with non-homoscedasticity is to squeeze out the effect of unequal variances with transformations of the X values. A number of transformations can be made (Dixon (1970) 17-19).

One way to examine the data for unequal variance is to plot the residuals of the regression for each independent variable against the dependent variable. A residual is the difference between the Y actually measured in the experiment and the Y value calculated during the regression. The calculated Y value is the appropriate point on the best-fit line drawn by the regression through all the data points. If the residuals for a given variable show a marked tendency to scatter in a particular pattern, then transformations of the data are probably required to assure homoscedasticity.

The regression program used for calculations in this thesis (see Section 3.6) could produce the required residual plots on demand. Examination of the plots of residuals for the major independent variables showed no distinct tendency in the scatter. Although there was a tendency for values to cluster at the low end of the x-axis where the independent variables (clue types) had values of 1 or 2, this effect was primarily due to the sparseness of high-valued observations. This was due to the fact that clues had a tendency to occur once, or twice, but seldom six or eight times in a single document. Of course, this meant that more data was available on the low end of the scale. The higher values seemed to be randomly scattered thoughout their ranges. For this reason, transformations of the original data were not necessary to preserve homoscedasticity.

3.3 The Correlation Coefficient

The correlation coefficient, R, is a measure of the strength of the linear relationship between the index terms assigned and the textual clues in the documents.

If the distributions of X and Y are similar, then R may take on any value from an extreme low value of -1 to an extreme high value of +1 (Hays (1963) 510). When the distributions of X and Y are very dissimilar, these extremes can shrink considerably (see Carroll (1961)). We would expect our X and Y distributions to be very similar. Most of the values of these two variables will be zero; a middling number of observations will have low values (one indexer assigns, or a clue occurs once in a document); fewer will have mid-range values (several indexers assign the same term, the same clue occurs several times); very few observations will have high values (almost all indexers agree to assign, a particular clue occurs many times in the document). An inspection of Figures 5.07 and 5.09 bears out this expectation. The indexers in Figure 5.07 have a tendency to make unique assignments; terms assigned by many indexers occur infrequently. The same distribution is evident in the totals of Figure 5.09. A particular clue type is usually a unique occurrance in a document. Since the distributions of X and Y in our data are very similar, R has a -1 to +1 range.

A +1 value of R means that the X and Y variables are perfectly positively correlated. In other words, Y varies in the same way and in the same direction as X because the possible values of X and Y lie on a straight line with a positive slope. If R has a value of -1, then X and Y are perfectly negatively correlated. This means that possible values of X and Y lie on a straight line with a negative slope. Between these two extremes, R can be zero. This means that X and Y are uncorrelated or linearly unassociated with each other. Two completely random phenomena exhibit a correlation coefficient of zero.

A correlation of +1, however, does not mean that there is a causal relationship between X and Y, nor does a correlation of zero mean that X and Y are statistically independent. We are simply observing that X and Y vary in a particular fashion, we are not saying why this variation occurs.

It should be noted that it is always possible to make R equal to 1 by increasing the number of independent variables to equal the number of observations made. As long as the number of variables (clue types) remains low in comparison to the number of observations, there is no danger of forcing the value of R to one. Thus, our ratio of 61 clue types to 6379 observations will not prejudice the value of R.

Recall from Section 3.2 that we have been using summed squared deviations as a measure of the best fit regression line. Again using summed squared deviations, the total variance exhibited by the data is equal to the summed squared deviations of the actual Y's from the average Y. This assumes that we merely averaged all the data. In fact, however, we are positing a linear relationship between X and Y, so the deviations we have not been able to explain by the regression equation are the summed squared deviations of the actual Y's from the Y's predicted by the regression equation. The explained variance is then the summed square of the difference between the Y's computed by the regression and the average Y. If we divide the explained variance by the total variance, then we have a measure of the amount of variance accounted for by the regression, or a measure of the "goodness" of the regression. This statistic is:

$$R^2 = \text{explained variance} / \text{total variance} \qquad (3.03)$$

and it is expressed as a percentage. In fact, it is the percentage of variance accounted for by the regression. Note that when R is either +1 or -1, R^2 is also one and that when R is almost zero, R^2 is also almost zero.

For our purposes, then, R^2 is the percentage of indexing accounted for by the regression. As far as the regression model is concerned, it is the percentage of indexing behavior which can be accounted for by the use of textual clues.

We would like to say how good the correlation coefficient is as an estimator of the true correlation coefficient. Any value of R may be transformed to a new variable, Z, in the following way (see Edwards (1967) 248-250 or Hays (1963) 530-531).

$$Z = (\ln (1 + R) - \ln (1 - R))/2. \qquad (3.04)$$

Fisher (1921) has shown that the distribution of Z is very close to normal with a mean of zero and a standard deviation of 1 and that Z is independent of the sample size. The standard error of Z is:

$$S = 1/\sqrt{(N - n - 1)}. \qquad (3.05)$$

where N is the number of experimental events and n the number of variables. The correlation coefficient for the entire population therefore lies between an upper bound of (Z + S • K) and a lower bound of (Z - S • K) where K is the percentage cut-off point on the normal curve (for a 99% confidence interval, K = 2.58). These upper and lower bounds on Z may be transformed back into R values so that a confidence interval may be set around the correlation coefficient.

We will be comparing the correlation coefficients obtained from different experimental sub-samples and would like to test the significance of the difference between two correlation coefficients. The Z transformation also permits this kind of test (Edwards (1967) 250-252).

3.4 Relative Importance of Clues

The regression program discussed in Section 3.6 adds variables to the regression equation one at a time, giving information after each addition about the improvement to R^2 caused by each variable. It will therefore tell us how much each variable contributes to the final value of R^2.

It is the improvement in R^2 effected by a clue as it enters the regression equation which indicates its importance in accounting for the indexing (see Section 3.3). R^2 measures the sum of the direct and indirect effects of each variable. A full discussion of the relative importance of clue types in the best regression equation will be found in Section 6.3.

3.5 Prediction with the Regression Model

After the line described by Equation 3.02 has been determined for the sample, the values obtained for the B coefficients can be used to predict values of Y for new documents. Since the B coefficients describe a line which is the closest fit for the experimental points, this line is the best available predictor for new documents.

Let us assume we wish to index a new document with the prediction function of our regression equation. For each thesaurus descriptor to be considered by the model there will be a set of X values, n per descriptor. The B coefficients have already been calculated from the sample documents. To estimate the percentage of indexers who will assign the first descriptor, the appropriate B and X values are multiplied together and the terms summed to get the value of Y.

The arithmetic is simple enough; logic subjects the process to some restrictions, however. First, it would obviously not be profitable to use the equation if the correlation coefficient itself is not high. If only a small part of indexer behavior can be accounted for by the textual clues, then it doesn't make much sense to try to use the clues as a substitute for human indexing.

Secondly, even if the average R is high, there may be a group of documents or terms for which the R is quite low. Thus it is important to know just how well the equation predicts assignment for each of the sample documents. If the predicted Y values vary wildly from the actual Y values for the documents in the sample, the use of the regression equation for prediction is not reasonable.

Third, we must not forget that there may well be some uncontrollable variables or some peculiar characteristics of the sample document set or indexers which influence the way clues are used. It would not be fair to generalize, for example, from a single sample of documents about instruments and instrumentation to all documents in any technical field.

Fourth, it is quite possible that the predicted value of Y may not fit our practical notions of what makes sense. The values of Y for the sample lie between zero and one ($1 \geq Y \geq 0$) because they represent the proportion of indexers assigning the term. Since proportions may not be negative or greater than one, negative values of Y and values of Y greater than one cannot occur. It is possible that when the regression equation is used on new documents, some particular combination of X's will make the predicted value of Y for the new document lie outside the zero to one common-sense limits. Statistically, there is nothing wrong with a predicted $Y < 0$ or $Y > 1$. If this occurs, we simply correct a $Y < 0$ to a zero and a $Y > 1$ to a one.

3.6 Computer Program for Multiple Linear Regression

The regression calculations were done with a stepwise multiple regression program, BMD02R, available from the University of California at Los Angeles, Health Sciences Computing Facility. This program calculates a series of multiple linear regression equations. The program searches for the independent variable (the clue type) with the highest correlation with the dependent variable (percentage of indexers assigning). The regression equation is then calculated. The independent variable with the next-highest correlation with those already in the equation is then chosen and the regression equation recalculated. Each new equation adds one new variable to the calculations. Draper and Smith ((1966) 163-195) may be consulted for a discussion of various computational procedures for regressions including stepwise regression.

After each variable is added to the equation, the program prints the multiple correlation coefficient R, the coefficient of multiple determination R^2, the standard error of estimate, an analysis of variance table, the regression coefficient, the value of F and the standard error for each variable in the equation, and other useful statistical information. Scatter plots of the residuals of each independent variable against the dependent variable are also available. The method for obtaining this stepwise information, not ordinarily available, was suggested by Efroymson (1960). The program will accept a maximum of 80 variables and 9999 experimental events or cases. Complete documentation of the program may be found in Dixon ((1970) 233-257).

The regression program is written in Fortran IV (H level) and uses Assembly language subroutines. A regression run for about 6400 experimental events and 61 variables requires about an hour of cpu time and from 3 to 12 hours elapsed time on an IBM 360/65. Confidence intervals were calculated with an interactive mathematical system called APL/360.

Chapter Four

THE COMBINATORIAL MODEL

4. The Combinatorial Model

4.1 Reasons for the Combinatorial Model

The combinatorial model is intended to cover exhaustively a class of non-linear assignment rules of the type discussed in Section 2.5.3. This model does not assume a linear relationship nor does it make an assumption of normality, except in a Central Limit sense.

4.2 Types of Indexer-Model Agreement

Let us assume we have a black box model of indexer behavior. If this model is fed the textual clues from a document and a term from the thesaurus, it replies with a yes/no answer. Let us also assume we have a human indexer. If this indexer is given an index term from the thesaurus and is asked whether that term should be assigned to a document, he, too, can give a yes/no answer.

This leads to four types or cases of indexer-model agreement for the assignment of a particular index term to a particular document:

case 1) neither indexer nor model assigns term,
case 2) both indexer and model assign term,
case 3) indexer assigns term, model does not,
case 4) model assigns term, indexer does not.

If the model always agrees with the indexer (case 1 and case 2 only), then it will be a perfect predictor of the human. The greater the number of decisions of case 3 and case 4 type, the worse the model is as a predictor of the human.

Assume, for the moment, that we wish to test the ability of a single textual clue to predict a human indexer's performance. Further, assume that each sample document has been tested for the presence or absence of this clue for each of five possible thesaurus terms and that the human has also registered his yes/no decision. The results of this test can be summarized in the following way:

document and term	clue present?	indexer decision	case type
document 1 term 1	no	no	case 1
document 1 term 2	yes	yes	case 2
document 1 term 3	yes	no	case 4
document 1 term 4	no	yes	case 3
document 1 term 5	no	no	case 1
document 2 term 1	no	no	case 1
document 2 term 2	no	no	case 1
document 2 term 3	no	no	case 1
document 2 term 4	no	no	case 1
document 2 term 5	yes	yes	case 2

We can summarize this example as:

	case 1	case 2	case 3	case 4
document 1	2	1	1	1
document 2	4	1	0	0

The case 1 through 4 totals for each document indicate how accurately the model predicts the performance of the indexer based on a single textual clue. If the model agrees with the indexer all of the time, only case 1 and case 2 exist (as document 2 illustrates). If the model is less successful, then case 3 and case 4 conditions may also exist (as document 1 illustrates).

This example has a thesaurus of 5 terms. An increase in thesaurus size almost necessarily increases the number of case 1 occurrences (neither the model nor the indexer assigns) since the index set for a document is not a function of the thesaurus size and seldom contains more than five or ten terms. In a larger thesaurus, the overwhelming majority of case 1's completely swamp out the other cases. Such a preponderance of agreement leads to an arithmetically impressive model, but since the case 1 agreements carry almost no information and disguise the occurrences of the rest of the cases, they must be dropped from the model.

For document 1, then, the model correctly predicted 1 out of 3 non-trivial assignments (that is, non-case 1 assignments) and thus accounted for 33% of the indexer's performance with a single textual clue. For document 2, the model predicted 1 out of 1 non-trivial assignments, accounting for 100% of the indexer's performance. The non-trivial assignments are a measure of how well the model matches the indexer. The figure-of-merit for non-trivial assignments is calculated as:

$$\text{Figure-of-merit} = (\text{case } 2)/(\text{case } 2 + \text{case } 3 + \text{case } 4). \qquad (4.01)$$

This figure-of-merit is often called a "precision ratio" and is commonly used in document systems to quantify the success of the system in answering requests. Becker and Hayes ((1963) 370-372) point out that this measure "attaches no weight at all to agreement in 0's and is therefore only suitable where the proportion of 1's to 0's is low. It is the most obvious definition in those cases where, at any rate in principle, the columns are indefinitely long but the number of 1's in each is fixed or (statistically) limited" (page 371). The 0's of Becker and Hayes are our case 1; their 1's are our cases 2 through 4. Since our thesaurus is very large in comparison to the number of terms in a single document's index set, our situation is an appropriate one in which to use Equation 4.01.

Although this measure is an appropriate one for us, there is nothing in the combinatorial model preventing the use of a different figure-of-merit. In fact, a second figure-of-merit is employed in Section 6.2 for the comparison of the Boolean combinatorial model and the regression model. This second figure-of-merit includes case 1's:

$$\text{Fraction of all predictions modelled} = (\text{case } 1 + \text{case } 2) \,/\, (\text{all cases}). \qquad (4.02)$$

The following papers may be consulted for more extensive discussions of measures of nearness or coefficients of association: Kuhns (1965), Jones and Curtice (1967).

We have calculated the success of the single-clue model in predicting a single indexer's behavior for two documents. The calculation can be repeated for any number of documents. We could then report on the average success of the single-clue model in predicting that particular indexer's behavior by averaging the scores obtained for each of the individual documents.

We could also obtain this average figure-of-merit for each of a group of indexers. We could then average the averages to obtain an over-all figure of merit to summarize the success of the single-clue model in predicting group indexing behavior.

Similar calculations could be made for any other clue to be used in the model. We could then compare the over-all figure of merit for each of these clue types to say which ones did a better job of predicting human indexing behavior.

4.3 Boolean Combinations of Clues

The Boolean combinations covered in this section are intended to test a group of textual clues and a class of selection rules in an exhaustive fashion. Obviously, other combinatorial rules can be imagined and tried out, and other types of textual clues could also be investigated. If this first, exhaustive trial is successful, additional refinements might be worthwhile.

We will be using two types of Boolean operators to represent two types of indexing behavior. If the indexer behaves as if both of two clue types are required to motivate assignment, then AND behavior is displayed. For example, suppose we consider the thesaurus term 'radiation counters'. Suppose an indexer assigns the term only if the word 'radiation' and the word 'counters' are both present (but not necessarily contiguous) in the document. The indexer is saying 'radiation' AND 'counters' lead to the assignment of 'radiation counters'. This is AND behavior. Of course, AND behavior may combine more than two clue types in a single expression.

If the indexer behaves as if either of two clues could motivate him, then he is exhibiting OR behavior. For example, suppose the thesaurus term is 'vertical takeoff aircraft' and the indexer assigns the term whenever either the term itself or a used-for reference, 'convertiplanes', or both occurs in the document. Either 'vertical takeoff aircraft' OR 'convertiplanes' leads to the assignment of 'vertical takeoff aircraft'. This is OR behavior. OR behavior may also combine more than two clue types in a single ORed expression.

The exhaustive Boolean combination proceeds in the following manner. First, each single-clue type is tested. Each of the documents in the sample is tested for the presence of each clue type for each term in the thesaurus. The presence or absence of a clue type for a thesaurus term is recorded in a yes/no indicator. Then, as discussed in the previous section, each indexer's behavior is compared against the single-clue model and the results summarized by the average figure-of-merit discussed there. The individual indexer figure-of-merit for a particular clue type is averaged to yield an over-all figure-of-merit for each single clue. This information is saved for later use in the model.

Next, the yes/no indicators for every pair of clues are ANDed together. This produces a new yes/no indicator for the presence or absence of that ANDed pair of clue types for each thesaurus term. Each indexer's behavior is compared against the two-clue-ANDed-model and the results summarized by an average figure-of-merit. The individual-indexer performance figures for a particular clue type are averaged to yield an over-all figure-of-merit for each pair of ANDed clues. This information is also retained for later use in the combinatorial model.

Next, the procedures described above for use on all single clues and all pairs of clues are repeated for all triplets and quadruplets of clues and the information saved for later use. Four was chosen as a maximum number for this ANDing step because it appeared to be well beyond the complexity humans might use in clue selection.

One would expect that much ANDing of single clues would eventually produce a yes/no indicator consisting of nothing but no's or zeros. These clue combinations cannot help in the modelling since there are no terms which both the indexer and the model agree to assign (that is, there are no case 2's). These unfruitful clue combinations are dropped from further consideration.

At this stage in the procedure, we have produced and saved all possible ANDed combinations of single, double, etc. clues which might have some value later on in the model. In order to have some value, the combinations must have shown evidence of at least one thesaurus term for one document for which the ANDed clue combination correctly predicted that the indexer would assign the term.

The next step is to test all possible ORed combinations of the clues from the AND step. Each of the ANDed clues will be ORed with all the other ANDed clues. After each trial ORing takes place, the over-all figure-of-merit is calculated for the new ORed combination under test. After pairs of

ANDed clues have been ORed together, triplets of ANDed clues are ORed, then quadruplets, etc.

One would expect that much ORing of the ANDed clues would also eventually produce lower over-all figures-of-merit since the incidence of indexer-model agreement (that is, case 2, as discussed in the previous section) can only increase to a maximum of five or ten for each document, while the incidence of indexer model disagreement (that is, case 3 and case 4 as discussed in the previous section) could increase considerably beyond this. If ORing produces new trial clue combinations with a decreased over-all figure of merit, further ORing of these clues is terminated.

The end result of this sequence of ANDing and ORing is a group of equations of the following form:

$$(C1) \text{ OR } (C2 \text{ AND } C3) \text{ OR } (C4 \text{ AND } C5 \text{ AND } C6) \text{ OR } \ldots \quad (4.03)$$

where C1 through C6 are arbitrary clue types. Each ANDed element in the equation may be composed of a single clue, or pairs, triplets or quadruplets of clues ANDed together. Any number of ANDed elements may be combined with OR operators. Hence the equations, and each term within them, may be variable.

Each of these Boolean equations is associated with an over-all figure-of-merit which summarizes how well that particular equation predicts the average performance of the group of indexers. Because of the sequence of ANDing and ORing operations, these remaining Boolean equations are guaranteed to have the highest figure-of-merit. This is therefore the set of equations which most accurately predicts how the indexers behaved on the average. It is the best set of models of human indexing behavior which we can build with the specified procedure.

Ideally, we wish to obtain the simplest model which will predict accurately how humans index. We are therefore looking for the equation with the highest figure-of-merit and with the least number of ANDed and ORed terms.

4.4 Statistical Tests of the Combinatorial Model

Statistical tests of the significance of the combinatorial model are much less complex than those for the regression model. The over-all figure-of-merit for the highest ranking Boolean equation quantifies the amount of human indexing accounted for by the textual clues. The figure-of-merit for each of the equations is simply an average of all indexer behavior, over all documents, for all thesaurus terms in the sample. To be able to make statements about the entire population of indexers, documents and thesaurus terms, from this sample, we use the Central Limit Theorem (Hays (1963) 238-244) to obtain a normally distributed population. For the Boolean equation with the highest average figure-of-merit, we know how well the equation predicts the average indexing for each document-term pair. If random scores chosen from this large sample are averaged, a normal distribution is produced. From this normal distribution the standard deviation of the sample may be calculated. The confidence interval for whatever confidence coefficient we choose can then be obtained.

One of the major points of interest is a comparison of the scientist-indexers against the librarian-indexers to determine which group is most accurately represented by the textual clue model. If the figures-of-merit are calculated for the indexing of the scientist-indexer group alone, the Central Limit Theorem provides standard deviations just as it did above for the total indexer group. The calculations can be repeated for the librarian-indexer group.

The relative importance of the textual clues is immediately available from an observation of the equations themselves. It is of interest to know which clues are most frequently used in the ANDed and ORed equations.

The predictive properties of the Boolean equations are straightforward. A new document is tested for the existence of each of the clue types. These binary values are plugged into the Boolean equation. The decision on the assignment of each thesaurus term is "yes" if the Boolean equation returns a value of one, and "no" if a zero is obtained.

As with the regression model, we must use caution when applying the model predictively. The Boolean equation is not a universal automatic indexer just becaase it may account for the human indexing behavior on a sample of documents. There might well be special circumstances affecting our group of documents and indexers which render the model inaccurate when used on a radically different sample.

4.5 Computer Programs for the Combinatorial Model

The computer programs discussed in this section were written in PL/I and run on an IBM 360/65. Assembly language subroutines were used to generate random numbers and to count the number of ones in a bit string.

After the comparison of the document words with the thesaurus, (see Section 5.2.5) there were a total of 12,440 clue vectors. Of these, 6061 recorded no matches with the thesaurus and no indexer assignments for that particular index term. In other words, the entire clue and indexer vector was zero. ANDing and ORing of these all-zero vectors would not have affected the Boolean model, so they were eliminated from further processing as far as this model was concerned. From the remaining 6379 non-zero vectors, 2048 vectors were chosen randomly with the random number generator proposed by Lewis, Goodman and Miller (1969). This particular sample size was chosen because the IBM 360 machines can perform Boolean operations on a bit string of length 2048 in a single machine instruction.

The master vector for each of these 2048 observations was then read into core and organized in an array. This array was 60 bits wide (one bit for each clue type) and 2048 bits high (one bit for each observation on the sample). The array was then transposed so that it could be efficiently handled in later Boolean operations. The same procedure was followed with the indexer array. Recorded in the master vector set was information about whether each indexer assigned a particular term, or did not assign it. Each indexer's choice of terms then could be represented as an array one bit wide(one bit to indicate whether the term was assigned or not) and 2048 bits high (one bit for each observation in the sample). Since there were twelve indexers, the array was actually 12 bits wide. This array was also transposed so that it could be compared efficiently with the clue array.

The ANDing program then processed the clue and indexer array in the following manner. The clue array, now 2048 bits wide by 60 high, was read into core. The indexer array, now 2048 bits wide by 12 high, was also read into core. The program then tested the first clue against all twelve indexers. It did this by ANDing the clue vector with the first indexer's vector and counting the number of one bits in the 2048 bit string. Counting was done with an Assembly language subroutine suggested by Raduchel (1970). The number of one bits in the ANDed string equaled the number of observations in which the clue vector agreed with the indexer - that is, the number of case 2's in the sample. This is the numerator of the figure-of-merit. The same first clue was then ORed with the same indexer vector, and the one bits in the ORed string counted. The number of one bits in this ORed string equaled the number of observations in which either the clue type or the indexer indicated a term should be assigned - that is, the number of case 2's plus case 3's plus case 4's in the sample. This is the denominator of the figure-of-merit. This sequence of ANDing and ORing a clue vector with the indexer vector was repeated for each of the twelve indexers. The resulting average figure-of-merit was then stored with the clue pattern and vector for later use in the ORing program. Of course, if the figure-of-merit was zero, the vector and the information about it were discarded.

After processing the first clue vector in this manner, the program then ANDed the first clue vector with the second and tested the resulting vector against the indexer vectors. It then tried ANDing in the third vector, and so forth. When the program had tried all possible ANDed combinations involving the first clue vector, it then moved on to the second. This ANDing sequence was chosen to minimize access time in core. The result of this processing was a total of 5572 ANDed vectors. The best of these vectors had a figure-of-merit of 0.11517.

The ORing programs were organized in a similar manner, except that there was not enough core storage or computer time to handle all 5572 ANDed vectors. For this reason, the best 300 ANDed vectors were processed one at a time against the other ANDed vectors. ORing of pairs of ANDed vectors produced a total of 45,150 ORed vectors with a high figure-of-merit of 0.15051. ORing continued, one stage at a time, until a maximum of eight ANDed clues had been ORed together. The vector with the highest figure-of-merit was separated by sorting and is discussed in Chapter 6.

The confidence interval around the best figure-of-merit was obtained by taking random selections of 32 observations from the 2048 observations in the final best vector. The individual figures-of-merit for each of these smaller groups were calculated and the results used to compute the confidence interval.

Chapter Five

EXPERIMENTAL PROCEDURES AND SAMPLES

5. Experimental Procedures and Samples

5.1 The Documents and Indexers

A group of scientists and engineers (see Section 2.6) with experience in the field of instrumentation was available to serve as scientist-indexer subjects. To cater to their field of specialization, all documents indexed by any of the following terms were selected from the 1969 subject index of U.S. Government Research and Development Reports (USGRDR): acoustic measuring instruments, aircraft instruments, astronomical instruments, charge measuring instruments, electrically powered instruments, electric measuring instruments, meteorological instruments, optical measuring instruments, pneumatic instruments, radiation measuring instruments, recording instruments, spacecraft instruments, strain measuring instruments, surveying instruments, temperature measuring instruments, thermal measuring instruments, time measuring instruments, voltage measuring instruments. These terms are the set of descriptors with 'instruments' as the last word with two exceptions, surgical instruments and musical instruments, which were not included because they fell outside the usual range of instrument subject expertise for the individuals involved.

The 1969 USGRDR indexes contained 78 documents indexed under the above terms. These documents were arranged in ascending order by the report number. A random number table was used to select twenty documents to serve as a test sample. The complete information for each of these twenty documents was then keypunched directly from the USGRDR entry (see Section 5.2.1 for details). Only the title and abstract were used in the experiments discussed here. Hereafter, the word "document" means only the title and abstract of the document as those titles and abstracts appear in USGRDR.

Two groups indexed each of the twenty documents. The first group consisted of the six librarian-indexers and the second, the six scientist-indexers. Each indexer was given the same set of materials from which to work. This set consisted of 1) the titles and abstracts of each of the documents to be indexed in a standard printed format, 2) indexing instructions and 3) the Engineers Joint Council (EJC) Thesaurus of Engineering and Scientific Terms (1967). The standard document format was produced by a computer program which arranged each document on the page so no words were broken at the end of a line. Some standard information was printed at the bottom of each page. The documents were printed on alternate pages so the indexer could see only a single document at a time. See Figure 5.01 for a reduced copy of one page of this printout.

The instructions to the indexers are reproduced in Figure 5.02. A page from the EJC Thesaurus is reproduced in Figure 5.03. The terms chosen by the indexer for each document were keypunched and a computer program then collected the individual index sets for each of the documents and for each thesaurus phrase assigned. This program provided the "terms-assigned" information for the programs discussed in Section 5.2.

5.2 Clue Counting Procedures

The problem of finding, identifying and counting particular types of clues in natural language text is common to both of the indexing models used in this thesis. When even moderate numbers of clues must be located, the task becomes much too tedious to be done accurately by hand. For this reason, computer programs were written to find and count each clue type. All of the computer programs discussed in this section were written in PL/I and run on an IBM 360/65.

5.2.1 Keypunching

Each document in the sample was keypunched, proof-read and corrected. In general, the text was keypunched exactly as printed. Exceptions to this rule were caused by the limited keypunch character set:

1 If the document contained a character not on the keypunch, the word for that character was substituted. This rule was very seldom needed.

2 When words were broken with a hyphen over the end of a justified line of printed text, the hyphen was dropped and the word "glued together" again in the keypunching.

3 Subscripts and superscripts were keypunched on the line with the text.

4 All lower case letters in the printed text were keypunched as upper-case characters.

Figure 5.04 shows the original printed version of one of the documents. The machine-printed version of this document is shown in Figure 5.01.

A program was written to isolate each word in the running text. This program considered a word to be any sequence of the alphabetical characters (A...Z) unbroken by a non-alphabetic character (0123...9,:;/,etc.). Since none of the thesaurus terms contained non-alphabetic characters, this procedure did not discard any potential matches. Each of the single words was written on a sequential file with information on the document being processed, the location of that word in the document (title or abstract) and the relative position of the word in the document (counting the first word in the document as one, the second word as two, etc.).

5.2.2 Reduction to Singular Form

The matching procedure detailed in later sections of this chapter considers singular and plural forms of a word to be equivalent. Each of the words isolated in the previous section was tested for the ending 'ies', 'es' or 's'. If a word ended in 'ies', this ending was changed to a 'y'; if the word ended in 's', the 's' was dropped; it the word ended in 'es' the ending was dropped after sibilants ('s', 'ss', 'c', 'sh', etc.). Exceptions to these general rules were programmed individually. For instance the singular forms of 'pulses' and 'mars' do not follow the regular rules and were therefore handled as exceptions.

Since the comparison had to be made between the document and the thesaurus, the same procedure was followed for the words from each of the thesaurus descriptors. Figure 5.05 shows the singular form of some words from the document in Figure 5.04.

5.2.3 Stemming

The root segment of each of the words was then found with the stemming algorithm suggested by Lovins (1968). This algorithm searches for the longest match in a list of endings ordered by length. If a match occurs, and if context-sensitive conditions associated with that ending are satisfied, the program strips the ending from the word. The resulting stem is then additionally transformed with recoding rules which handle spelling exceptions.

To minimize search time, the list of endings was hashed with the division method (see Lum, Yuen and Dodd (1971) for a comparison and review of various hashing techniques). A number stored at the hash location pointed into a separate table which resolved clashes and itemized the context-sensitive conditions to be satisfied for each of the endings. If the conditions were satisfied, the recoding procedures were invoked. The resulting stem was then paired with the original word in a record comprised of document number, location, and relative position. Figure 5.05 also shows the stemmed form of some words from the document in Figure 5.04. The appendix summarizes the additions and changes to Lovin's endings, conditions and recoding rules necessitated by the vocabulary in our sample.

5.2.4 Thesaurus Terms Used in the Models

The Engineers Joint Council Thesaurus contains 17,810 descriptors. Most of the thesaurus would have no matches with any sample document and would not be assigned by any of the indexers. Thus, most of the thesaurus could reasonably be expected to have an indexer and clue vector consisting entirely of zeros. These experimental points would be useless for this investigation. For this reason, the size of the thesaurus was reduced for processing in the following way. First all index terms assigned by any of the indexers to any of the documents were included in the thesaurus. There were 430 of these terms. This group of terms includes, for any particular document, all the clue vectors which have non-zero indexer values.

To include other vectors with guaranteed non-zero clue values in the vector, a sort and count was made of all the words in all the documents. Omitting function words such as 'as', 'a', 'the', the most frequently used words were used to search the complete EJC Thesaurus for descriptors containing these words. Descriptors containing these frequently used words were added to the first group of 430 descriptors. Note that this procedure forces the models to account for, not just the assignment of descriptors, but also the non-assignment of likely descriptors. This choice makes the model more conservative in ascribing machine-like behavior to the humans. The final mini-thesaurus contained 622 terms.

5.2.5 Document-Clue Matching Procedure

As discussed in Section 2.5, matching phrases, synonyms, words and roots in the thesaurus and in each document were counted to produce what we are calling a "clue vector". For each document-descriptor pair, this clue vector summarizes the number of times each clue type appears in the document.

Information on the number and types of clues existing in each document was obtained from a program which compared each descriptor in the mini-thesaurus against the words of each document in the sample. The program first hashed a document's words into core storage. A single thesaurus phrase was then read in. It was compared with the words of the document by hashing the thesaurus words and searching for matches with the hashed document words. If matches did occur, the clue vector for that document-thesaurus pair was updated with the appropriate information and the program then read in the next thesaurus phrase. After the entire mini-thesaurus has been compared with the first document, the hashing locations were cleared so that the next document's words could be processed. This program processed a total of 12,440 vectors for the documents in about 30 minutes.

5.3 Sub-Samples Tested

The result of the processing described in Section 5.2 is a set of 12,440 clue vectors, 622 clue vectors for each of the 20 documents in the sample. We will call this set the 'master set'. Of the 12,440 vectors, 6061 were completely zero in both indexer assignments and clues; 6379 were non-zero in at least one portion of the record.

Since the difference in indexing behavior between scientist-indexers and librarian-indexers is of considerable interest, two new sets of 12,440 clue vectors each were produced for these two groups of indexers. Each of the new vectors sets was based on the indexing done by the appropriate indexer group.

Several other subsets were taken. Since many of the studies in Chapter 2 considered only the terms assigned by the indexers, a subset of vectors was made by separating only those terms which were assigned by at least one of the indexers. These vectors should show greater evidence of "machine-like" indexing than the rest of the master set, if the effects noted in Chapter 2 hold. A second subset was made by separating only those terms assigned by two or more indexers.

It is also of interest to know how each document and indexer varies from the average. Information on the documents is obtained by processing the clue vectors for each document separately. Information on each indexer is obtained by re-running the entire model with clue vectors based only on the indexer in question. These runs will characterize individual documents and indexers in detail. They might, for instance, reveal a group of documents which are modelled extremely well, and a group which are not modelled sucessfully. Further inspection of these documents may help to explain the success or failure of the model. Five documents were selected randomly for individual processing: documents 1, 2, 6, 14 and 20. Four indexers were selected randomly for individual processing: 4, 6, 7, 11.

5.4 Statistics Describing the Documents and Indexers

To give the reader a feeling for the document sample, some numerical parameters summarizing the incidence of clues have been tabulated in this section. Figure 5.06 gives information on the length of the documents in the sample. Figure 5.07 lists the number of terms which were assigned by from one to twelve indexers. For example, on document 4, eleven of the indexers agreed one of the terms should be assigned, while there were 26 terms assigned by just one of the indexers. Figure 5.08 summarizes the number of times each clue type occurred in the entire document sample. Note that the number of clues occurring in the title were always less than the number occurring in the abstract. This is because the title was short in comparison to the abstract. Figure 5.09 gives the distribution density of all clue types in each of the sample documents. For instance, of the 37,320 possible clues for each document (622 thesaurus terms times 60 clue types) 508 clues appeared once each in document 1. However, 61 clues appeared four times each in document 1.

Figure 5.01 Document from the printout used by the indexers

DOCUMENT 4

TITLE: FLIGHT PROTOTYPE MODEL METEOR FLASH ANALYZER

DOCUMENT: A FLIGHT PROTOTYPE METEOR FLASH ANALYZER WITH A THREE CHANNEL RADIOMETER WAS DESIGNED, CONSTRUCTED AND TESTED. EACH CHANNEL HAS VIDEO OUTPUTS TO MEASURE THE INTENSITY VERSUS TIME VARIATION OF INDIVIDUAL METEOR FLASHES; AND THERE ARE A TOTAL OF 9 METEOR DATA CHANNELS FOR MAKING RELATED MEASUREMENTS. THE LONG WAVELENGTH (IRON) CHANNEL NEARLY COINCIDES WITH THE CONVENTIONAL SPECTRAL RANGE FOR PHOTOGRAPHIC METEORS, PROVIDING CORRELATION WITH GROUND BASED OBSERVATIONAL DATA. DETECTION SENSITIVITY FOR TERRESTRIAL METEORS IN THE IRON CHANNEL IS BACKGROUND RADIATION LIMITED; AND THIS APPEARS TO YIELD SUPERIOR SENSITIVITY FOR THE OPTICAL DETECTION OF METEORS IN WAVELENGTH BANDS BELOW THE OZONE LIMIT AT 0.30 MICRONS. FOR SATELLITE ALTITUDE OF NAUT M, DETECTOR FIELD OF VIEW OF 30 DEGREES, AND DETECTOR APERTURE DIAMETER OF 5 CM, LIMITING PHOTOGRAPHIC METEOR MAGNITUDE WAS +3.3, WITH AN INVERSE COUNT RATE OF 5.6 MIN/METEOR. AT 600 NAUT M, COUNT RATE IS EXPECTED TO BE 22 MIN/METEOR WITH A MAGNITUDE OF -1.1. SIGNIFICANTLY LARGER COUNT RATES ARE EXPECTED FOR THE MAGNESIUM AND SILICON CHANNELS.

DESCRIPTORS FOR THIS DOCUMENT:

1 ______________________________

2 ______________________________

3 ______________________________

4 ______________________________

5 ______________________________

6 ______________________________

7 ______________________________

8 ______________________________

9 ______________________________

10 ______________________________

Figure 5.02 Instructions for the indexers

INSTRUCTIONS

Imagine that you are a professional indexer for STAR (Scientific and Technical Aerospace Reports) or USGRDR (U.S. Government Research and Development Reports). Both of these indexing and abstracting journals are distributed internationally to engineers and scientists interested in current information in their fields.

On each page of the enclosed printout is a document. Below the document are numbered blank lines on which you are to record your choice of indexing terms for that document. The terms must be chosen from the enclosed EJC Thesaurus of Engineering and Scientific Terms. (If you are not familiar with this thesaurus, see its description following these instructions.) Space is provided for up to ten indexing terms. If you wish to assign more than ten terms to a document, simply write in the additional terms at the bottom of the page.

You may find it helpful to note the important subjects while reading over the document. You may use the space below the document for this purpose. The thesaurus can then be used to rephrase these subjects into the appropriate index terms on the numbered lines or below them.

Choose the most appropriate (applicable or useful) terms from the thesaurus for each document. Any number of terms may be assigned. Be as specific as possible in assigning terms. Remember you are indexing for engineers and scientists who will want to find these documents for their own research. The terms you assign should enable them to locate pertinent information quickly.

Please keep track of the time you spend indexing. Use the right-hand side of the printout page to record each time you begin indexing in hours and minutes, for instance: begin 4:32. When you are interrupted or have to quit, record the end time as: end 5:25. This job should be least imposing if you choose a time and a place permitting extended periods of concentration without disturbance.

In summary, you have two tasks:
1) Assign the best index terms to each document,
2) Keep track of all time spent indexing.

If you have any questions at any time, please call me collect at home (415-327-0727) or at work (408-227-7100 ext. 5435 or ext. 5611). Many thanks for your help.

Caryl

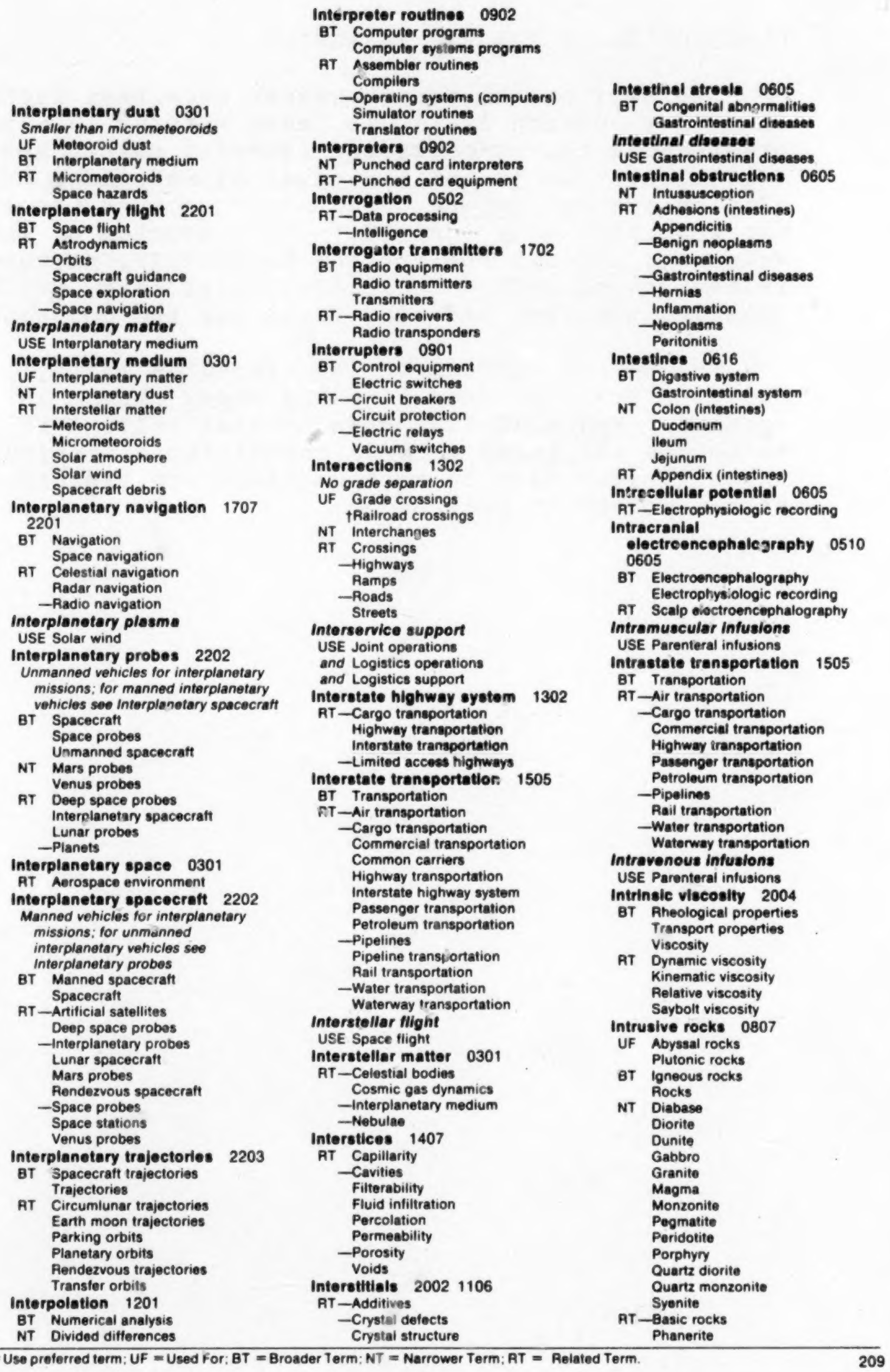
Figure 5.02 Instructions for the indexers (continued)

DESCRIPTION OF THE EJC THESAURUS

Two sections of the thesaurus have been marked with tabs. The first section lists all index phrases in alphabetical order character-by-character ignoring spaces and punctuation. (Note that this is not the usual alphabetical order. "Bands" appears before "Band saws" because the blank in the two word phrase is ignored.) This section of the thesaurus makes suggestions for broader terms (BT), narrower terms (NT) and related terms (RT). These additional terms may be useful in locating the best indexing terms for the document.

The second section of the thesaurus lists, in alphabetical order, every word used in every index phrase in the first section. You will find this section helpful if you would like to locate all index phrases containing a particular word. Abbreviations used in both sections are explained in footnotes at the bottom of each page.

Figure 5.03 Page from the Engineers Joint Council Thesaurus



Interplanetary dust 0301
Smaller than micrometeoroids
UF Meteoroid dust
BT Interplanetary medium
RT Micrometeoroids
Space hazards

Interplanetary flight 2201
BT Space flight
RT Astrodynamics
—Orbits
Spacecraft guidance
Space exploration
—Space navigation

Interplanetary matter
USE Interplanetary medium

Interplanetary medium 0301
UF Interplanetary matter
NT Interplanetary dust
RT Interstellar matter
—Meteoroids
Micrometeoroids
Solar atmosphere
Solar wind
Spacecraft debris

Interplanetary navigation 1707 2201
BT Navigation
Space navigation
RT Celestial navigation
Radar navigation
—Radio navigation

Interplanetary plasma
USE Solar wind

Interplanetary probes 2202
Unmanned vehicles for interplanetary missions; for manned interplanetary vehicles see Interplanetary spacecraft
BT Spacecraft
Space probes
Unmanned spacecraft
NT Mars probes
Venus probes
RT Deep space probes
Interplanetary spacecraft
Lunar probes
—Planets

Interplanetary space 0301
RT Aerospace environment

Interplanetary spacecraft 2202
Manned vehicles for interplanetary missions; for unmanned interplanetary vehicles see Interplanetary probes
BT Manned spacecraft
Spacecraft
RT—Artificial satellites
Deep space probes
—Interplanetary probes
Lunar spacecraft
Mars probes
Rendezvous spacecraft
—Space probes
Space stations
Venus probes

Interplanetary trajectories 2203
BT Spacecraft trajectories
Trajectories
RT Circumlunar trajectories
Earth moon trajectories
Parking orbits
Planetary orbits
Rendezvous trajectories
Transfer orbits

Interpolation 1201
BT Numerical analysis
NT Divided differences

Interpreter routines 0902
BT Computer programs
Computer systems programs
RT Assembler routines
Compilers
—Operating systems (computers)
Simulator routines
Translator routines

Interpreters 0902
NT Punched card interpreters
RT—Punched card equipment

Interrogation 0502
RT—Data processing
—Intelligence

Interrogator transmitters 1702
BT Radio equipment
Radio transmitters
Transmitters
RT—Radio receivers
Radio transponders

Interrupters 0901
BT Control equipment
Electric switches
RT—Circuit breakers
Circuit protection
—Electric relays
Vacuum switches

Intersections 1302
No grade separation
UF Grade crossings
†Railroad crossings
NT Interchanges
RT Crossings
—Highways
Ramps
—Roads
Streets

Interservice support
USE Joint operations
and Logistics operations
and Logistics support

Interstate highway system 1302
RT—Cargo transportation
Highway transportation
Interstate transportation
—Limited access highways

Interstate transportation 1505
BT Transportation
RT—Air transportation
—Cargo transportation
Commercial transportation
Common carriers
Highway transportation
Interstate highway system
Passenger transportation
Petroleum transportation
—Pipelines
Pipeline transportation
Rail transportation
—Water transportation
Waterway transportation

Interstellar flight
USE Space flight

Interstellar matter 0301
RT—Celestial bodies
Cosmic gas dynamics
—Interplanetary medium
—Nebulae

Interstices 1407
RT Capillarity
—Cavities
Filterability
Fluid infiltration
Percolation
Permeability
—Porosity
Voids

Interstitials 2002 1106
RT—Additives
—Crystal defects
Crystal structure

Intestinal atresia 0605
BT Congenital abnormalities
Gastrointestinal diseases

Intestinal diseases
USE Gastrointestinal diseases

Intestinal obstructions 0605
NT Intussusception
RT Adhesions (intestines)
Appendicitis
—Benign neoplasms
Constipation
—Gastrointestinal diseases
—Hernias
Inflammation
—Neoplasms
Peritonitis

Intestines 0616
BT Digestive system
Gastrointestinal system
NT Colon (intestines)
Duodenum
Ileum
Jejunum
RT Appendix (intestines)

Intracellular potential 0605
RT—Electrophysiologic recording

Intracranial electroencephalography 0510 0605
BT Electroencephalography
Electrophysiologic recording
RT Scalp electroencephalography

Intramuscular infusions
USE Parenteral infusions

Intrastate transportation 1505
BT Transportation
RT—Air transportation
—Cargo transportation
Commercial transportation
Highway transportation
Passenger transportation
Petroleum transportation
—Pipelines
Rail transportation
—Water transportation
Waterway transportation

Intravenous infusions
USE Parenteral infusions

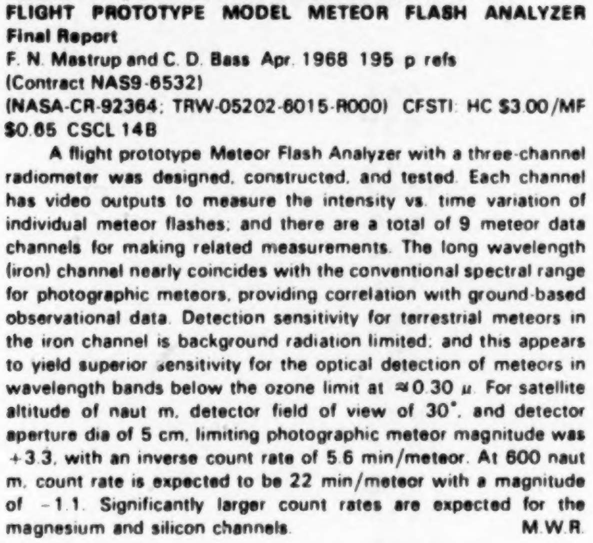
Intrinsic viscosity 2004
BT Rheological properties
Transport properties
Viscosity
RT Dynamic viscosity
Kinematic viscosity
Relative viscosity
Saybolt viscosity

Intrusive rocks 0807
UF Abyssal rocks
Plutonic rocks
BT Igneous rocks
Rocks
NT Diabase
Diorite
Dunite
Gabbro
Granite
Magma
Monzonite
Pegmatite
Peridotite
Porphyry
Quartz diorite
Quartz monzonite
Syenite
RT—Basic rocks
Phanerite

USE = Use preferred term; UF = Used For; BT = Broader Term; NT = Narrower Term; RT = Related Term.

209

Figure 5.04 Printed version of the document in Figure 5.01



N68-38439*# TRW Systems Group, Redondo Beach, Calif.
FLIGHT PROTOTYPE MODEL METEOR FLASH ANALYZER
Final Report
F. N. Mastrup and C. D. Bass Apr. 1968 195 p refs
(Contract NAS9-6532)
(NASA-CR-92364; TRW-05202-6015-R000) CFSTI: HC $3.00/MF
$0.65 CSCL 14B

A flight prototype Meteor Flash Analyzer with a three-channel radiometer was designed, constructed, and tested. Each channel has video outputs to measure the intensity vs. time variation of individual meteor flashes; and there are a total of 9 meteor data channels for making related measurements. The long wavelength (iron) channel nearly coincides with the conventional spectral range for photographic meteors, providing correlation with ground-based observational data. Detection sensitivity for terrestrial meteors in the iron channel is background radiation limited; and this appears to yield superior sensitivity for the optical detection of meteors in wavelength bands below the ozone limit at ≈0.30 μ. For satellite altitude of naut m, detector field of view of 30°, and detector aperture dia of 5 cm, limiting photographic meteor magnitude was +3.3, with an inverse count rate of 5.6 min/meteor. At 600 naut m, count rate is expected to be 22 min/meteor with a magnitude of −1.1. Significantly larger count rates are expected for the magnesium and silicon channels. M.W.R.

Figure 5.05 Singular and stemmed forms of some words from the document in Figure 5.04

singular form	stemmed form
flight	flight
prototype	prototyp
model	model
meteor	meteor
flash	flash
analyzer	analys
designed	design
constructed	construc
variation	vari
measurement	measur
photographic	photograph
providing	provid
observational	observ
sensitivity	sensit
detection	detect
detector	detect

Figure 5.06 Length of the sample documents

Document	Number of characters in document	Number of words in document
01	957	139
02	663	89
03	931	136
04	1164	166
05	1187	179
06	999	140
07	1208	169
08	606	86
09	1062	154
10	267	33
11	852	127
12	1078	153
13	1022	136
14	442	62
15	494	62
16	1288	207
17	1240	184
18	967	134
19	800	116
20	580	84
Total	7807	2556
Average	890	128

Figure 5.07 Distribution of Indexing Consensus on Sample

Average number of indexers assigning a term: 2.31

Standard deviation: 2.23

	Number of Indexers Assigning											
Document	1	2	3	4	5	6	7	8	9	10	11	12
1	17	6	3	1	1	0	1	1	0	0	0	0
2	11	6	0	0	0	2	0	0	0	0	0	1
3	23	5	5	2	2	0	1	0	0	1	0	0
4	26	9	0	0	0	0	1	0	0	0	1	0
5	10	11	2	1	3	0	1	0	0	1	0	0
6	14	7	3	1	0	1	0	0	0	1	2	0
7	16	4	0	0	0	1	0	2	1	2	0	0
8	13	3	3	0	4	0	1	0	0	0	0	0
9	22	8	4	2	2	1	0	0	0	1	0	0
10	13	4	2	1	0	2	0	0	0	0	0	1
11	15	7	2	1	2	0	0	0	0	1	0	0
12	25	4	5	4	3	2	1	3	0	0	1	0
13	21	14	3	1	4	0	0	0	1	1	0	0
14	13	1	2	1	0	1	0	0	1	0	1	0
15	17	3	0	4	0	1	0	0	1	0	0	0
16	18	7	2	0	1	1	0	0	1	0	0	0
17	14	4	5	1	1	0	2	0	0	0	0	0
18	18	3	1	1	3	1	1	1	0	0	0	0
19	14	4	3	2	0	1	0	0	0	2	0	0
20	7	2	2	1	1	1	0	0	0	0	1	0
Totals	327	112	47	24	27	15	9	7	5	10	6	2

Figure 5.08 Number of Clue Cccurrences in Entire Sample

Type of Match	MN	ST	US	PR	NR	RL	Total
3T	0	0	0	0	0	0	0
3A	1	2	0	0	0	0	3
2T	5	6	0	0	2	1	14
2A	61	73	17	10	54	105	320
HT	324	448	149	243	600	1492	3256
HA	2216	3246	1141	1845	4509	11758	24715
M2T	215	297	84	180	285	988	2049
M2A	1632	2245	634	1087	2593	7352	15543
M1T	18	31	10	25	38	68	190
M1A	220	283	139	220	498	808	2168
Totals	4692	6631	2174	3610	8579	22572	48258

(See Figure 6.01 for an explanation of acronyms.)

Figure 5.09 Distribution of All Clue Occurrences in Documents

	Number of Occurrences/Document for All Clues									
Document	1	2	3	4	5	6	7	8	9	10+
1	508	254	73	61	16	27	16	12	3	21
2	465	222	77	46	18	10	2	4	4	10
3	575	256	188	117	23	43	15	27	7	36
4	681	232	59	29	16	16	3	11	18	16
5	486	178	113	51	28	20	16	17	6	52
6	617	214	90	52	84	35	44	48	13	85
7	365	103	70	12	10	79	4	2	2	22
8	327	93	60	13	15	17	5	1	3	6
9	460	264	34	127	11	34	4	15	4	24
10	349	88	19	3	8	0	1	2	0	8
11	598	397	89	54	13	22	6	11	1	18
12	693	372	87	71	24	18	10	12	7	27
13	936	183	296	78	40	64	13	11	25	38
14	312	40	14	10	0	0	1	4	2	2
15	602	231	42	24	64	12	4	5	2	23
16	358	110	52	23	49	4	212	22	7	69
17	532	132	56	25	5	88	9	5	2	24
18	452	241	76	46	28	24	12	9	1	13
19	746	183	149	51	22	27	7	6	6	29
20	449	152	103	26	9	15	3	4	3	16
Totals	10511	3945	1747	919	483	555	387	228	116	539

Chapter Six

CONCLUSIONS

6. Conclusions

6.1 Introduction

This chapter discusses the results of the models described in Chapters 3 and 4. To simplify the following discussion and to save repeating long names, each of the clues has been assigned a brief descriptive name. These acronyms are listed in Figure 6.01 (pages 122 and 123) together with a fuller description of the clue.

6.2 Evidence For and Against Machine-Like Indexing

6.2.1 Results of the Multiple Linear Regression

Information about the major regression runs is summarized below.

All Indexers

- 6379 experimental events
- 47 clue types with correlation greater than .0001 with dependent variable
- multiple correlation coefficient (R): 0.5386
- square of correlation coefficient (R^2): 0.2901
- 99% confidence interval for R: .5153 to .5611

Librarian Indexers

- 6379 experimental events
- 45 clue types with correlation greater than .0001
- multiple correlation coefficient (R): 0.5364
- square of correlation coefficient (R^2): 0.2877
- 99% confidence interval for R: .5130 to .5590

Engineer and Scientist Indexers

- 6379 experimental events
- 46 clue types with correlation greater than .0001
- multiple correlation coefficient (R): 0.4674
- square of the correlation coefficient (R^2): 0.2184
- 99% confidence interval for R: .4418 to .4923

Perhaps the most dramatic result is that none of the samples taken shows either a very strong or a very weak correlation between the descriptors assigned and the documents. At least for our samples, linear regression accounts for about thirty percent of the indexing assignments.

As expected, the librarians' indexing could be predicted more accurately from the clues than could the indexing of the engineers and scientists. The difference was significant at the 99% level. The inexperience of the engineers and scientists with indexing and with the thesaurus may have made them much more dependent upon word-for-word matches between the descriptors and the document than otherwise might have been the case. Hence our results are probably conservative. Differences between librarians and engineers or scientists might be more pronounced under other experimental conditions.

It is difficult to compare the results of our multiple linear regression model with results obtained from previous studies because of a number of differences in the studies.

First, there is presently not enough data on how the subject content of the sample documents affects the results. The documents used in our sample were in some instances highly technical discussions of rather specific engineering problems. The subject field of our documents compares most closely with the studies done by Slamecka and Zunde (1963).

Unfortunately, they made only a cursory examination of the data from the viewpoint of machine-like indexing.

Secondly, there is the problem of the size of the sample of indexers. We used a total of 12 indexers, six librarians and six engineers or scientists. All twelve indexed each of the sample documents. No previous study had such a large group of indexers.

Thirdly, most previous studies did not account for the non-assignment of index terms as discussed in Chapter 2. The effect of looking only at assigned terms is demonstrated by re-running the regression on only those experimental events which have an indexer value above zero. Two runs were made. In the first, a term had to be assigned by at least one indexer to be included in the regression; in the second, a term had to be assigned by at least two indexers to be included. Information about these two sub-samples is given below.

At least one indexer assigned each term
591 experimental events
46 clue types with correlation greater than .0001
multiple correlation coefficient (R): 0.5486
square of the correlation coefficient (R^2): 0.3009
95% confidence interval for R: .4896 to .6026

At least two indexers assigned each term

264 experimental events

41 clue types with correlation greater than .0001

multiple correlation coefficient (R): 0.5585

square of correlation coefficient (R^2): 0.3120

95% confidence interval for R: .4694 to .6363

As expected, R^2 increases as more of the indexers agree to assign a particular term but the results are not striking or significant. Because of the small number of experimental events in which a majority of indexers agreed, and because of the large number of independent variables, the confidence intervals for these coefficients is considerably larger than for the full sample size.

Fourth, although all of the above regressions tested the effect of sixty possible clue types on the indexers, they still could account for only about a third of the variance in the indexing. This is in contrast to earlier studies which tock only one or two clue types into account, but which did not consider the non-assignment of index terms. The effect of the small number of clue types would, in general, be to decrease the correlation between the clue types and the indexing. The inclusion of only assigned index terms would tend to have the opposite effect. This is probably why our numerical results are very roughly comparable to some studies done with fewer clues and based on assigned terms.

Lastly, it is also possible that there is a theoretical maximum to the amount of indexing which can be matched with document words. For example, we can imagine a thesaurus which was specifically designed for a particular group of documents. This imaginary thesaurus might contain only words and phrases abstracted directly from the documents themselves. In this case, there is little opportunity for the indexer to assign a term not already in the document. We could also imagine a second thesaurus which made it a rule never to use a document word or phrase as a descriptor. Although such a thesaurus would probably be very difficult to compile, it would guarantee that there was no correlation between the index terms assigned and the words or phrases in the documents.

The EJC Thesaurus obviously lies somewhere between these two extremes. It is quite possible, therefore, that there could only be a certain number of matches between the terms and the document words simply because of the nature of the documents and the thesaurus. The extent of this theoretical limitation on the amount of the potential match between the documents and the thesaurus might account for differences between results obtained by different experimenters.

6.2.2 Results of the Combinatorial Model

The combinatorial model is based on Boolean combinations of the sixty clue types. Details of the best Boolean equation (produced according to the procedure described in Section 4.3) are given below:

Best Boolean equation
Sample size: 2048
Figure-of-merit (non-trivial assignments): .1611
Standard deviation of figure-of-merit: .0353
Fraction of all predictions modelled (trivial and non-trivial assignments): .6821
Case 2's: 125
Case 3's and 4's: 651

In terms of programming, the Boolean combinatorial model was time consuming and difficult. Despite careful program design and coding, it took over an hour of cpu time on an IBM 360/65 to OR 40,000 pairs of vectors, calculate a figure-of-merit for each, and write the results on tape. Similar run times were required for each stage of ANDing and ORing. Because of these very long computer runs, the combinatorial model is not exhaustive. Instead, as discussed in Chapter 4, the best 300 vectors from the previous stage were used to calculate vectors for the succeeding stage.

Another limitation of the combinatorial model was the practical limitation on the recording of count information for each type of clue in a document. Equation 2.02 is based on a zero/not-zero decision. Thus there is no difference in the binary record between a clue type which occurred just once and one which occurred many times. Once again, this is a practical decision necessitated by limited computer time. The lack of clue count information, however, makes this model less rich than the regression model.

A limitation on the sample size for the combinatorial model was also made for computational reasons. However, the particular sample taken was verified with the regression model by running that model with both the limited and the full data. The regression coefficient for the smaller sample of 2048 was 0.5387, just 0.0001 larger than it was for the sample of 6397. The limited sample of 2048, therefore, is representative of the full sample size of 6397.

The figure-of-merit based on non-trivial assignments (see Section 4.2 for definitions of "non-trivial" and "figure-of-merit") is quite low. There were only a few case 2's in the best vector and a number of case 3's and 4's. As discussed in Section 4.4, the standard deviation of the figure-of-merit was calculated by making use of the Central Limit Theorem. Sixty-four samples of thirty-two each were chosen at random from the large sample of 2048. This produced the standard deviation of the figure-of-merit of .0353.

The goodness of the model can also be judged in terms of the number of case 1's and 2's divided by the total number of cases. This is the second figure-of-merit (called "fraction of all predictions modelled") introduced in Equation 4.02 in Section 4.2 This means that out of 2048 possible indexing decisions, the combinatorial model duplicated 68% of the indexers' decisions. This method of calculating this number is more comparable with the regression model and will be discussed in Section 6.2.3.

It is unfortunate that more computer time was not available. It would have been interesting to repeat the combinatorial model for the sub-samples used with the regression model and to compare the results. As can be seen from the discussion of the relative importance of clues in Section 6.3, the combinatorial model has a more direct interpretation of indexer behavior than does the regression model. Perhaps further refinement of the programming and the elimination of less valuable clue types may make it possible to include count information and larger sample sizes in a future version of the combinatorial model.

6.2.3 Comparison of the Results of the Models

Primarily because the Boolean model did not make use of the clue count information in the documents, and because "best" was defined differently in the two models, there is no simple, direct comparison between the two models. To make the figures from the two models somewhat more comparable, a second figure-of-merit was calculated for the Boolean model.

This is the number recorded above and discussed in Section 4.2 as "fraction of all predictions modelled".

Both the combinatorial and the regression models were run on the same sample of 2048. The combinatorial model accounted for 68% of all indexer decisions. That is, of the 2048 decisions, there were 1397 decisions in which the model correctly predicted what the indexers assigned. The model assigned when the indexers did, and did not assign when they didn't. For the same sample, the regression model had an R^2 of 0.3009. In other words, approximately 30% of the variance in indexing could be accounted for by the regression. In view of the different ways in which these two percentages were calculated, the amount of indexing accounted for by the two models may be comparable. The lower percentage obtained from the regression is probably due to the linearity assumed by this model.

In Section 2.5.3 we discussed the assignment rules tested in the Boolean and regression models and pointed out that in some special cases the two models are equivalent. Each of the four Boolean equations was tested for this equivalence (that is, linear separability) with the Biswas (1971) method. None can be realized with a single threshold element. Hence there is no direct mathematical equivalence between the two models.

Neither of the models performed well enough to be useful as a substitute for human indexing. A discussion of prediction with these models has, therefore, been omitted. However, the values of A and of the B's for the first five steps of the regression are tabulated in Figure 6.05. Notice that as each new variable is added the previous values of the constant and of the B's change. The regression is adjusted at each stage for the best fit, changing the coefficients for the variables at each stage. As an example, let us take the fifth step in the regression. All the variables are positively related to Y. The higher the number of occurrences of each of these five clue types, the more likely the indexer to assign the descriptor. On the average, the number of indexers assigning a term increases by one unit for each three additional occurrences of a two-word main term in the abstract, by two units for each additional occurrence of a stemmed header in the title, and so forth.

The constant and coefficients for the full regression equation are tabulated in the right-hand column of Figure 6.02. Since the regression equation accounts for such a small percentage of indexer performance, this tabulation is not of much practical value.

In summary, then, at least for this sample and this rather large group of indexers, we cannot model very much of technical indexing with either a regression model or a Boolean combinatorial model. Until we know more about differences between technical fields, the effect of the thesaurus on the indexing, etc., it is invalid to argue that indexers in general act in a mechanical manner.

6.3 Relative Importance of the Clue Types

We have some specific evidence about the relative importance of the clue types from each of the models. In addition, we can compare the clue types important in the engineer/scientist regression with the clues important in the librarian regression. (See Section 2.5.1 for a definition of each clue type and Figure 6.01 for a table of all clue types and their acronyms.)

We can make no statements about the value of some clues in predicting indexing assignments because these clues did not occur in the sample. There were no title occurrences of any three-word descriptors, or of use, broader, narrower or related three-word terms in the abstract. Nor were there any two-word title occurrences of use or broader terms in the document sample. (See Figure 5.08 for a summary of clue occurrences in each of the documents.) Note that these would be the document-thesaurus matches least likely to occur in any sample because the match criterion was the most stringent (two and three word matches in the title and three word matches in the abstract). Note also that a high frequency of a clue type does not mean that the clue is necessarily important in predicting indexer behavior. There do, however, have to be enough cccurrences of a clue type to make that clue of practical value in the prediction.

Figure 6.02, 6.03, and 6.04 list the clues which have a correlation greater than .0001 with the dependent variable for each of the three major samples: all indexers, librarian indexers, and engineer/scientist indexers. These figures also show the R^2 at each step and the increase in R^2 caused by the addition of each variable to the regression. Although the order of importance varied from sample to sample, the same clue types tended to be at the top of the list.

In all runs, a match of a two-word descriptor in the abstract was the most important of the clues. This single clue accounted for 63 to 75% of the final value of R. Other clues consistently occurred in the top group of all three regression runs. They were modifier2 of use references in the abstract, two-word use references in the abstract, modifier1 of broader terms in the title and modifier2 of the stemmed term in the abstract. Although main and stem two-word terms in the title, main three-word terms in the abstract and modifier1 use references in the title were also in the top group, they are less important because they occurred infrequently in the sample. Thus main entries, use references, modifiers1, modifiers2, and two and three word phrases are most important clues in predicting indexer assignments.

There are also several clues which rank high in the regression for all indexers, but which have quite different rankings when the engineer and librarian regressions are compared although no test of statistical significance was made. These clues are stemmed header terms in the title, the main term header in the abstract and the use term header in the abstract. The stemmed header in the title is rated low by the librarians and high by the scientists; the main term and use term headers in the abstract are rated high by the scientists and low and mid-range respectively by the librarians. Apparently the header word of a descriptor is treated differently by the librarians and scientists. Note that there are no header clues in the top group agreed upon by all indexers as important.

Let us contrast the clue ranking of the regression model with that given by the Boolean model. The best four Boolean equations are given below.

(MN 2A AND ST 2A) OR (MN HA AND MN M2T) OR (US M2T AND MN HA)

(MN 2A AND ST 2A) OR (MN HA AND MN M2T) OR (US M2T AND ST HA)

(MN 2A AND ST 2A) OR (MN HA AND MN M2T) OR (US M2T AND MN HA AND US M2A)

(MN 2A AND ST 2A) OR (MN HA AND MN M2T) OR (US M2T AND ST HA AND US M2A)

where:

MN 2A	is a two-word main term in the abstract
ST 2A	is a two-word stem in the abstract
MN HA	is a main term header in the abstract
MN M2T	is a main term modifier2 in the title
US M2T	is a use reference modifier2 in the title
US M2A	is a use reference modifier2 in the abstract

Each of the top equations contains the same two ANDed terms plus a third term which is variable. The clue types mentioned in all of the equations are main entries, stems or use references. Narrower, broader and related terms do not serve as good clues in the Boolean equations.

The first of the ANDed expressions is a very simple requirement. If the descriptor has two words, then it must appear, as a phrase, in the abstract of the document if the descriptor is to be assigned. (Recall that there were no three-word abstract or two-word title occurrences (see Figure 5.08) so that these clues did not occur in high enough numbers to be represented in the final equation.) Of course, the stem of any term occurs whenever the term itself occurs by the clue definition rules in Section 2.5.1. The result is consistent with the results of the regression model where two-word main terms in the abstract account for a large part of the final value of the regression coefficient.

The second ANDed expression represents a second way to recognize a two-word phrase. The header for the descriptor and the modifier2 for that descriptor must be present in the document. Since most descriptors are two-word phrases, this is simply another way of saying that the words of the phrase must be present in the document.

The third ANDed expression is variable, but always contains US M2T and either MN HA or ST HA. In two of the equations, US M2A is an additional clue. Again MN HA and ST

HA are almost equivalent, so this last ANDed expression becomes: (MN HA and US M2T (and sometimes US M2A)). An inspection of the use references and the main terms when these clues occur shows that in many cases the modifier2 for the main term was the same as the modifier2 for the use-for reference. For example: 'optical instruments' use 'optical measurements'. Hence this ANDed expression once again reduces to: find the two-word descriptor phrase in the document.

In summary, two-word phrases account for the largest amount of indexing behavior. Some potentially valuable clue lengths, such as three-word terms, do not occur at all or in large enough numbers to make possible a decision about their value. Main, use and stemmed terms are the most important thesaurus relations. In general, broader, narrower and related terms from the thesaurus are not very useful in accounting for indexing behavior. Header terms are rated differently by the two sub-samples, but are not important for the entire sample. Finally, no generalizations can be made about the relative importance of title and abstract clues in accounting for indexer performance.

6.4 Individual Documents and Indexers

The regression coefficients obtained for five randomly selected documents in the sample were most interesting. The pertinent information is summarized below.

	Document Number				
	1	2	6	14	20
experimental events:	622	622	622	622	622
signif. clue types:	37	29	36	25	40
correlation coeff.:	.6412	.8574	.7228	.7868	.9084
R^2:	.4111	.7351	.5224	.6190	.8252
lower 99% conf. int.:	.5760	.8274	.6695	.7440	.8885
upper 99% conf. int.:	.6983	.8825	.7687	.8232	.9249
most important clues:	MN HA	MN 2A	RL 2A	US 2A	MN 2T
	US 2A	US M1A	RL M1T	MN HA	US 2A
	MN M2A	MN M2T	MN M1T	MN HT	NR M1T
	MN 2A	BR M1T	MN 2A	BR HA	ST HA

The significance of a clue type depends upon its contribution to the total regression. A clue type was considered significant if it had a correlation of at least .0001 with the dependent variable.

Note that the most important clue types and the correlation coefficients vary widely from document to document. In general, the correlation coefficient is considerably higher for an individual document than it is for the sample as a whole. This means that the regression coefficient for all the documents is very much a compromise. The compromise lowers the overall coefficient because clues which work well on some documents don't work well on others. As we noted in Section 3.5, this fact decreases the predictive value of this model.

Separate regression runs were made on the indexing of four of the subjects. Some details of these runs are summarized below.

	Indexer Number			
		6	7	11
experimental events:	6379	6379	6379	6379
signif. clue types:	43	48	47	48
correlation coeff.:	.3487	.4646	.3120	.2799
R^2:	.1216	.2158	.0974	.0783
lower 99% conf. int.:	.3200	.4389	.2825	.2499
upper 99% conf. int.:	.3768	.4896	.3409	.3094

For each indexer MN 2A was the top clue, accounting for 70, 78, 64 and 59% respectively of the correlation coefficient for the four indexers. After this clue, however, there were substantial differences among the top group of clues in the regression. The uniform use of MN 2A as the most important clue probably accounts for the top ranking of that clue in the over-all regression runs.

The fact that each indexer exhibited a low correlation coefficient as an individual, while single documents had high correlation coefficients indicates that there tends to be a common reaction to a single document, but that averaging across documents tends to decrease the correlation coefficient because the average is a compromise in clue styles among the documents. Individual indexers tend to be less predictable than an indexer group because one person's idiosyncrasies are not averaged with another's idiosyncrasies.

6.5 Some Suggestions for Further Research

This dissertation concentrated on textual clues to the exclusion of other types of clues (such as syntactic). Further investigation of other types of clues might help explain the existence of distinctive clue styles in individual documents. When these styles can be recognized from information about the document itself, we will have a better understanding of how an indexer goes about indexing.

Although the Boolean model is of much interest, a shortage of computer time prevented its full development. Further research might uncover practical improvements to speed up or to simplify the ANDing and ORing programs so that a more extensive development of this model could be made.

Our research was limited to an exploration of twenty documents in the rather narrow subject field of instrumentation. Since variations in indexing style are to be expected across subject fields, it would be interesting to build similar models in other subject fields and to compare the results.

Neither the regression nor the Boolean combinatorial models could be considered very accurate models of human indexing. However, as can be seen from Figure 5.07, humans themselves don't agree as to which index terms should be assigned. Inaccurate though these models are, it would be interesting to use them predictively and to ask humans how they rated the indexing derived from this mechanical source. Perhaps these models produce indexing no worse than a human's.

There is an implied "theory of the indexer" in this study which assumes that the indexer can be modelled by some combination of textual clues. The object of the investigation was to find out which clues were most important and how much of the indexing they accounted for. This is a very elementary theory of how indexing proceeds. A future study could begin to lay down a much more sophisticated theory of the indexer with some of the evidence available from this dissertation. For instance, two-word terms seem to be the most dependable for purposes of prediction. Suppose we start with a model to predict just two-word terms. We might say that if the term under consideration is a two-word term, then if that term, or if a stemmed version of that term is in the document, then the term should be assigned. Further elaboration of this simple flowchart model could be tested against the actual index terms assigned until some reasonable fit occurred. We could then test this flowchart model against other indexers to learn how accurate and complete it is.

Figure 6.01 Clue Types Used in the Two Models and the Acronyms Used for Them

number	acronym	description
1	MN 3T	three-word main descriptor entry in title
2	MN 3A	three-word main entry in abstract
3	MN 2T	two-word main entry in title
4	MN 2A	two-word main entry in abstract
5	MN HT	header word in title
6	MN HA	header word in abstract
7	MN M2T	modifier word of main entry in title
8	MN M2A	modifier2 word of main entry in abstract
9	MN M1T	modifier1 word of main entry in title
10	MN M1A	modifier1 word of main entry in abstract
11	ST 3T	three-word stem descriptor in title
12	ST 3A	three-word stem descriptor in abstract
13	ST 2T	two-word stem in title
14	ST 2A	two-word stem in abstract
15	ST HT	header stem in title
16	ST HA	header stem in abstract
17	ST M2T	modifier2 word of stem in title
18	ST M2A	modifier2 word of stem in abstract
19	ST M1T	modifier1 word of stem in title
20	ST M1A	modifier1 word of stem in abstract
21	US 3T	three-word use reference in title
22	US 3A	three-word use reference in abstract
23	US 2T	two-word use reference in title
24	US 2A	two-word use reference in abstract
25	US HT	header of use reference in title
26	US HA	header of use reference in abstract
27	US M2T	modifier2 word of use reference in title
28	US M2A	modifier2 word of use reference in abstract
29	US M1T	modifier1 word of use reference in title
30	US M1A	modifier1 word of use reference in abstract
31	BR 3T	three-word broader term in title
32	BR 3A	three-word broader term in abstract
33	BR 2T	two-word broader term in title
34	BR 2A	two-word broader term in abstract
35	BR HT	header word of broader term in title
36	BR HA	header word of broader term in abstract
37	BR M2T	modifier2 word of broader term in title
38	BR M2A	modifier2 word of broader term in abstract
39	BR M1T	modifier1 word of broader term in title
40	BR M1A	modifier1 word of broader term in abstract

Figure 6.01 Clue Types Used in the Two Models and the Acronyms Used for Them (continued)

number	acronym	description
41	NR 3T	three-word narrower term in title
42	NR 3A	three-word narrower term in abstract
43	NR 2T	two-word narrower term in title
44	NR 2A	two-word narrower term in abstract
45	NR HT	header of narrower term in title
46	NR HA	header of narrower term in abstract
47	NR M2T	modifier2 of narrower term in title
48	NR M2A	modifier2 of narrower term in abstract
49	NR M1T	modifier1 of narrower term in title
50	NR M1A	modifier1 of narrower term in abstract
51	RL 3T	three-word related term in title
52	RL 3A	three-word related term in abstract
53	RL 2T	two-word related term in title
54	RL 2A	two-word related term in abstract
55	RL HT	header word of related term in title
56	RL HA	header word of related term in abstract
57	RL M2T	modifier2 word of related term in title
58	RL M2A	modifier2 word of related term in abstract
59	RL M1T	modifier1 word of related term in title
60	RL M1A	modifier1 word of related term in abstract

Figure 6.02 Relative Importance of Clue Types for All Indexers

clue	R	R^2	Increase in R^2	B Coefficient of Full Equation	
MN 2A	.3863	.1492	.1492	1.81124	(A=0.05118)
US M2A	.4347	.1890	.0397	0.08572	
ST HT	.4565	.2084	.0194	0.18623	
US 2A	.4771	.2277	.0193	2.69215	
US M1T	.4901	.2402	.0126	1.87261	
ST M2T	.4998	.2498	.0096	0.09849	
MN HA	.5060	.2560	.0062	0.04049	
BR M1T	.5097	.2598	.0038	0.84427	
US HA	.5126	.2628	.0030	0.05867	
ST 2T	.5153	.2655	.0027	6.79555	
MN 2T	.5197	.2701	.0046	-5.80765	
ST M2A	.5224	.2729	.0028	0.04035	
MN 3A	.5247	.2753	.0025	5.45449	
US M2T	.5269	.2776	.0023	0.34074	
US M1A	.5288	.2796	.0020	0.10880	
RL 2A	.5302	.2811	.0014	0.25544	
ST 2A	.5312	.2822	.0011	0.72929	
BR HT	.5321	.2831	.0010	-0.13666	
MN M1T	.5329	.2840	.0009	0.31201	
RL M2T	.5336	.2847	.0006	0.06040	
RL 2T	.5342	.2854	.0007	-2.23822	
NR HA	.5348	.2860	.0006	0.00686	
MN HT	.5353	.2865	.0005	0.22254	
ST HA	.5357	.2870	.0005	0.03465	
MN M2T	.5361	.2874	.0004	0.14794	
US HT	.5364	.2877	.0004	-0.13123	
RL HT	.5367	.2881	.0004	0.03457	
MN M1A	.5370	.2884	.0003	0.07221	
NR M1T	.5372	.2886	.0002	-0.23186	
NR HT	.5374	.2888	.0002	-0.01565	
RL M1T	.5376	.2890	.0002	0.12045	
NR 2T	.5378	.2892	.0002	-0.79571	
RL HA	.5379	.2894	.0002	-0.00342	
ST 3A	.5380	.2895	.0001	-1.62292	
NR M1A	.5382	.2897	.0002	-0.23186	
BR M2A	.5383	.2898	.0001	0.01967	
BR HA	.5384	.2899	.0001	-0.01159	
MN M2A	.5384	.2899	.0001	0.02115	
RL M2A	.5385	.2900	.0001	-0.00404	
BR M2T	.5385	.2900	.0001	-0.02564	
NR M2T	.5385	.2900	.0000	0.01266	
BR 2A	.5386	.2900	.0000	-0.10282	
ST M1T	.5386	.2901	.0000	0.09310	
RL M1A	.5386	.2901	.0000	0.00524	
BR M1A	.5386	.2901	.0000	-0.00881	
ST M1A	.5386	.2901	.0000	-0.01367	
NR 2A	.5386	.2901	.0000	-0.01191	

Figure 6.03 Relative Importance of Clue Types for Librarian Indexers

clue	R	R^2	Increase in R^2
MN 2A	.4068	.1655	.1655
US 2A	.4558	.2077	.0423
ST HA	.4735	.2242	.0165
US M2A	.4889	.2390	.0148
ST M2A	.4978	.2478	.0087
US M1T	.5046	.2546	.0068
ST 2T	.5097	.2598	.0053
MN 2T	.5139	.2641	.0042
MN HT	.5180	.2683	.0042
ST 2A	.5207	.2712	.0029
BR M1T	.5226	.2731	.0019
RL M2T	.5242	.2748	.0017
MN 3A	.5257	.2764	.0016
MN M1T	.5271	.2778	.0014
US M1A	.5283	.2791	.0013
ST M2T	.5293	.2801	.0010
US HA	.5302	.2812	.0010
BR HT	.5310	.2820	.0008
NR M1T	.5316	.2826	.0006
NR M2A	.5323	.2833	.0007
US M2T	.5327	.2838	.0005
US HT	.5332	.2843	.0005
MN M2A	.5336	.2847	.0004
BR M2A	.5340	.2851	.0004
RL 2A	.5343	.2855	.0004
RL 2T	.5347	.2859	.0004
BR HA	.5350	.2862	.0003
NR HA	.5352	.2864	.0002
NR 2A	.5354	.2866	.0002
NR HT	.5356	.2868	.0002
RL HT	.5357	.2870	.0001
RL HA	.5358	.2871	.0001
ST 3A	.5359	.2872	.0001
MN HA	.5360	.2873	.0001
ST HT	.5361	.2874	.0001
NR M2T	.5362	.2875	.0001
RL M1A	.5362	.2875	.0001
RL M2A	.5363	.2876	.0001
MN M1A	.5363	.2877	.0001
NR M1A	.5364	.2877	.0000
NR 2T	.5364	.2877	.0000
RL M1T	.5364	.2877	.0000
MN M2T	.5364	.2877	.0000
BR M1A	.5364	.2877	.0000
BR M2T	.5364	.2877	.0000

Figure 6.04 Relative Importance of Clue Types for Engineer and Scientist Indexers

clue	R	R²	Increase in R²
MN 2A	.2926	.0856	.0856
US M2A	.3506	.1229	.0373
ST HT	.3808	.1450	.0221
US M1T	.3998	.1598	.014[illegible]
MN M2T	.4109	.1688	.0090
US HA	.4214	.1776	.0088
US 2A	.4301	.1850	.0074
BR M1T	.4362	.1903	.0053
MN HA	.4399	.1935	.0032
US M2T	.4432	.1964	.0030
MN 3A	.4464	.1993	.0028
RL 2A	.4489	.2015	.0023
ST 2T	.4506	.2031	.0015
MN 2T	.4546	.2067	.0036
ST M2A	.4564	.2083	.0016
MN M1A	.4579	.2097	.0014
BR HT	.4594	.2111	.0014
RL M1T	.4605	.2120	.0010
RL HT	.4614	.2129	.0008
RL 2T	.4623	.2137	.0008
NR 2T	.4631	.2145	.0008
NR HA	.4639	.2152	.0007
US M1A	.4645	.2157	.0006
RL M2T	.4649	.2161	.0004
NR 2A	.4653	.2165	.0004
NR M2T	.4655	.2167	.0002
NR M2A	.4658	.2169	.0003
NR M1A	.4659	.2171	.0001
ST 3A	.4662	.2173	.0003
NR HT	.4664	.2175	.0002
MN HT	.4666	.2177	.0002
NR M1T	.4667	.2178	.0001
US HT	.4668	.2179	.0001
BR M2T	.4669	.2180	.0001
RL M2A	.4670	.2181	.0001
BR 2A	.4671	.2182	.0001
ST M1T	.4671	.2182	.0001
RL HA	.4672	.2183	.0001
BR M1A	.4672	.2183	.0001
MN M2A	.4673	.2184	.0000
RL M1A	.4673	.2184	.0000
ST HA	.4673	.2184	.0000
BR HA	.4673	.2184	.0000
ST M2T	.4674	.2184	.0000
MN M1T	.4674	.2184	.0000
ST M1A	.4674	.2184	.0000

Figure 6.05 The First Five Regression Equations

	Steps or Number of Variables in Regression				
Variable	1	2	3	4	5
constant	0.18468	0.15769	0.12279	0.12152	0.12027
MN 2A	3.09661	3.01652	2.96815	2.93191	2.94239
US M2A		0.27927	0.27808	0.22456	0.19035
ST HT			0.50526	0.50823	0.50695
US 2A				2.52391	2.64097
US M1T					2.76123

Appendix

CHANGES TO LOVIN'S STEMMING PROCEDURES

Appendix. Changes to Iovin's Stemming Procedures

This appendix summarizes only the changes and additions to the stems, codes and rules proposed by Lovin (1968). The original paper should be consulted for a complete description of the procedures and tables. These changes were required by the cabulary of the documents and thesaurus used in this thesis. The effect of each proposed change to Lovin's procedures was tested on Brown's "Normal and Reverse English Word List" (1963) to guarantee that the intended change was not a parochial one.

The procedure used in Section 5.2.3 for stemming document and thesaurus words is dependent upon a table of stems, a set of condition codes and a group of recoding rules. A word to be stemmed is compared with the table of stem endings. The object is to obtain the longest possible match between the end of a word and an ending in the table. With each ending in the table is an associated "condition code". This code specifies the conditions to be met for that particular stem. An ending is rejected if the conditions for that ending are not met. If the ending passes the condition code test, the remaining stem is subjected to the recoding rules to standardize spelling variations.

The following three tables give the changes and additions to the endings, condition codes and recoding rules used by Lovin.

Additions and Changes to Endings and Condition Codes:

ending	cond.code	ending	cond.code
ationship	B	oides	B
mentation	EE	ology	C
ications	G	aged	B
ological	C	ents	C
icantly	A	ered	DD
ination	D	ison	H
ionable	Q	ists	D
ionless	Q	ites	AA
alized	BB	ment	EE
atures	E	oide	B
earity	Y	oids	B
ements	B	ying	C
erized	H	ers	DD
inants	B	est	O
mental	EE	ety	O
ologic	C	ons	S
atics	B	ors	T
icals	A	er	DD
ivity	C	ly	C
mets	EE	y	J

Additions and Changes to the Condition Codes:

E	not after 'e' unless 'gr' precedes 'e'
H	only after 't', 'll' or 'r'
L	do not remove after 'u','x','s' unless 's' follows 'o' and minimum stem length is 3
R	minimum stem length is 3 and remove only after 'n' or 'r'
V	remove only after 'c' or 'r'
DD	remove only after 'd','z','t','r','h','w','g','l', except after 'met'
EE	do not remove after 'u' or 'e'

Additions and Changes to Recoding Rules:

5a	change 'ript' to 'rib'
15	change 'ex' to 'ec' except after 'l'
24	change 'end' to 'ens' except after 's' or 'm'
31	change 'ert' to 'ers' except after 'v' or 'p'
32	change 'et' to 'es' except after 'n' or 'k'
35	change 'mart' to 'mar'
36	change 'ary' to 'ari'

BIBLIOGRAPHY

Artandi, Susan: Automatic book indexing by computer. American Documentation 15:4 (1964 October) 250-257.

Artandi, Susan: Computer indexing of medical articles - Project MEDICO. Journal of Documentation 25:3 (1969 September) 214-223.

Baxendale, P.B.: Machine-made index for technical literature - an experiment. IBM Journal of Research and Development 2:4 (1958 October) 354-361.

Baxendale, Phyllis: An empirical model for computer indexing. In: Third Institute on Information Storage and Retrieval, Machine Indexing: Progress and Problems, 13-17 February 1961. American University, Washington, D.C. (1962) 207-218.

Baxendale, P.B. and E.C. Clarke: Documentation for an economical program for the limited parsing of English: Lexicon, grammar, and flowcharts. IBM, San Jose, Calif. (1966 August 16) Research Report RJ-386.

Becker, Joseph and Robert M. Hayes: Information storage and retrieval: tools, elements, theories. Wiley (1963).

Bernier, Charles L.: Indexing process evaluation. American Documentation 16:4 (1965 October) 323-328.

Biswas, N.N.: Testing and realization of threshold functions by the canonical composition matrix. Paper submitted to the IEEE Transactions of Electronic Computers (1971). Department of Electrical Engineering, St. Louis University, St. Louis, Missouri.

Bloomfield, Masse: Simulated machine indexing, Parts I through IV. Special Libraries 57:3 (1966 March) 167-171; 57:4 (1966 April) 232-235; 57:5 (1966 May-June) 323-326; 57:6 (1966 July-August) 400-403.

Bottle, Robert T.: The information content of titles in the engineering literature. IEEE Transactions on Engineering Writing and Speech 13;2 (1970 September) 41-45.

Brown, A.F. (compiler): Normal and reverse English word list. Prepared at the University of Pennsylvania under a contract with the Air Force Office of Scientific Research. (AF 49 (638)-1042), 8 volumes (1963).

Carroll, J.B.: The nature of data, or how to choose a correlation coefficient. Psychometrika 26 (1961) 347-372.

Clarke, D.C. and R.E. Wall: An economical program for the limited parsing of English. AFIPS Conference Proceedings, Volume 27, Part 1, 1965 Fall Joint Computer Conference, 307-316.

Damerau, Fred J.: An experiment in automatic indexing. American Documentation 16:4 (1965 October) 283-289.

Dennis, Sally F.: The construction of a thesaurus automatically from a sample of text. In: M.E. Stevens (ed.): Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, 17-19 March 1964. U.S. Government Printing Office, Washington, D.C. (1965 December 15) National Bureau of Standards Miscellaneous Publication 269, 61-148.

Dennis, Sally F.: The design and testing of a fully automatic indexing-searching system for documents consisting of expository text. In: George Schector (ed.): Information Retrieval, A Critical Review. Based on the Third Annual National Colloquium on Information Retrieval, 12-13 May 1966, Philadelphia, Pa. Thompson Book Company, Washington, D.C. (1967) 67-94.

Dixon, W.J. (ed.): BMD Biomedical computer programs. BMD02R, Stepwise regression. University of California Press, Berkeley (1970) 233-257.

Doyle, L.B.: Library science in the computer age. System Development Corporation, Santa Monica, Calif. (1959 December 17) Report SP-141.

Draper, N.R. and H. Smith: Applied regression analysis. Wiley (1966).

Edmundson, H.P. and R.E. Wyllys; Automatic abstracting and indexing - survey and recommendations. Communications of the ACM 4:5 (1961 May) 226-234.

Edwards, Allen L.: Statistical methods, second edition. Holt, Rinehart and Winston (1967).

Efroymson, M.A.: Multiple regression analysis. In: Anthony Ralston and Herbert S. Wilf (eds.): Mathematical methods for digital computers. Wiley (1960) 191-203.

Engineers Joint Council: Thesaurus of engineering and scientific terms, first edition. EJC, New York (1967 December).

Fangmeyer, Hermann and Gerhard Lustig: The EURATOM automatic indexing project. In: A. J. H. Morrell (ed.): Information Processing 1968, Proceedings of the IFIP Congress. North-Holland Publishing Company, Amsterdam (1969) 1310-1314.

Fels, Eberhard M. and Joan Jacobs: Linguistic statistics of legal indexing. University of Pittsburgh Law Review 24 (1963) 771-791.

Ferber, Robert: Market research. McGraw-Hill (1949).

Fisher, R.A.: On the "probable error" of a coefficient of correlation. Metron 1:4 (1921) 1-32.

Graves, Roy W. and Donald P. Helander: A feasibility study of automatic indexing and information retrieval. IEEE Transactions of Engineering Writing and Speech 13:2 (1970 September) 58-59.

Harris, Z.S.: Linguistic transformations for information retrieval. In: Proceedings of the International Conference on Scientific Information, 16-21 November 1958. National Academy of Sciences - National Research Council, Washington, D.C. (1959) v.2, 937-950.

Hays, William L.: Statistics for psychologists. Holt, Rinehart and Winston (1963).

Hooper, R.S.: Indexer consistency tests - origin, measurements, results and utilization. IBM Corporation, Bethesda, Md. (1965) Report TR-95-56.

Houston, Nona and Eugene Wall: The distribution of term usage in manipulative indexes. American Documentation 15:2 (1964 April) 105-114.

Jones, Paul E. and Robert M. Curtice: A framework for comparing term association measures. American Documentation 18:3 (1967 July) 153-161.

Jones, Paul. E., Vincent E. Giuliano and Robert. M. Curtice: Automatic language processing, Part I: Selected collection statistics and data analyses, Section II: Comparison of manual and machine selected vocabularies. American Data Processing, Detroit (1970) 31-45.

Klingbiel, Paul H.: Machine-aided indexing. Defense Documentation Center, Defense Supply Agency, Alexandria, Virginia (1969 June) DDC-TR-69-1, AD 696 200.

Klingbiel, Paul H.: Machine-aided indexing. Defense Documentation Center, Defense Supply Agency, Alexandria, Virginia (1971 March) DDC-TR-71-3, AD 721 875.

Kuhns, J.L.: The continuum of coefficients of association. In M.E. Stevens (ed.): Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, 17-19 March 1964. U.S. Government Printing Office, Washington, D.C. (1965 December 15) National Bureau of Standards Miscellaneous Publication 269, 33-39.

Lewis, P.A. W., A.S. Goodman and J.M. Miller: A pseudo-random number generator for the System/360. IBM System Journal 8:2 (1969) 136-146.

Lewis, P.M. and C.L. Coates: Threshold logic. Wiley (1967).

Lovins, Julie Beth: Development of a stemming algorithm. Mechanical Translation 11:1 and 2 (1968 March and June) 22-31.

Lum, V.Y., P.S.T. Yuen and M. Dodd: Key-to-address transformation techniques, a fundamental performance study on large existing formatted files. Communications of the ACM 14:4 (1971 April) 228-239.

Luhn, H. P.: A statistical approach to mechanized encoding and searching of library information. IBM Journal of Research and Development 1:4 (1957 October) 309-317.

Luhn, H. P.: The automatic creation of literature abstracts. IBM Journal of Research and Development 2:2 (1958 April) 159-165.

Luhn, H. P.: Potentialities of auto-encoding of scientific literature. IBM, Yorktown Heights, N.Y. (1959 May 15) Research Report RC-101.

MacMillan, Judith T. and Isaac D. Welt: A study of indexing procedures in a limited area of the medical sciences. American Documentation 12:1 (1961 January) 27-31.

Meyer-Uhlenried, K.H. and G. Lustig: Analysis, indexing and correlation of information. In: H.P. Luhn (ed.): American Documentation Institute, 26th Annual Meeting, Automation and Scientific Communication, 6 11 October 1963. ADI, Washington, D.C. (1963) part 2, 229.

Montgomery, Christine and Don R. Swanson: Machine like indexing by people. American Documentation 13:4 (1962 October) 359-366.

O'Connor, John: Some suggested mechanized indexing investigations which require no machines. American Documentation 12:3 (1961 July) 198-203.

O'Connor, John: Some remarks on mechanized indexing and some small-scale empirical results. In: Third Institute on Information Storage and Retrieval, Machine Indexing: Progress and problems, 13-17 February 1961. American University, Washington, D.C. (1962) 266-279.

O'Connor, John: Correlation of index headings and title words in three medical indexing systems. American Documentation 15:2 (1964 April) 96-104.

O'Connor, John: Automatic subject recognition in scientific papers: an empirical study. Journal of the ACM 12:4 (1965 October) 490-515.

Ostle, Bernard: Statistics in research, second edition. Iowa State University Press, Ames, Iowa (1963).

Penn, Carol A.: How an indexer thinks in describing information, in framing search questions and in conducting searches. Journal of Chemical Documentation 2:4 (1962 October) 220-224.

Raduchel, William J.: Efficient handling of binary data. Communications of the ACM 13:12 (1970 December) 758-759.

Rees, Alan M.: Relevancy and pertinency in indexing. American Documentation 13:1 (1962 January) 93-94.

Salton, G.: Automatic information organization and retrieval. McGraw-Hill (1968).

Salton, G.: Automatic text analysis. Science 168:3929 (1970 April 17) 335-343.

Scheffe, Henry: Analysis of variance. Wiley (1959).

Shapiro, Paul A., Irwin D.J. Bross, Roger L. Priore and Barbara B. Anderson: Information in natural languages. Journal of the American Medical Association 207:11 (1969 March 17) 2080-2084.

Simmons, Robert F.: Natural language question-answering systems: 1969. Communications of the ACM 13:1 (1970 January) 15-30.

Slamecka, V. and P. Zunde: Automatic subject indexing from textual condensations. In: H.P. Luhn (ed.): American Documentation Institute, 26th Annual Meeting, Automation and Scientific Communication, 6-11 October 1963. ADI, Washington, D.C. (1963) Part 2, 139-140.

St. Laurent, Mary Cuddy: A review of the literature of indexer consistency. University of Chicago Master's Dissertation (1966 June) PB 174 395.

Stevens, Mary Elizabeth: Automatic indexing, a state-of-the-art report. U.S. Government Printing Office, Washington, D.C. (Reissued with corrections, 1970 February). National Bureau of Standards Monograph No. 91.

Stone, Don Charles: Word statistics in the generation of semantic tools for information systems. University of Pennsylvania, Philadelphia (1967 December) AD 664 915.

Stone, Don C. and Morris Rubinoff: Statistical generation of a technical vocabulary. American Documentation 19:4 (1968 October) 411-412.

Swanson, Don R.: Searching natural language text by computer. Science 132:3434 (1960 October 21) 1099-1104.

Swanson, Don R.: Interrogating a computer in natural language. In: Cicely M. Popplewell (ed.): Information Processing 1962, Proceedings of the IFIP Congress. North-Holland Publishing Company, Amsterdam (1963) 288-293.

Swanson, Don R.: Automatic indexing and classification. Preprint, NATO Advanced Study Institute on Automatic Document Analysis, 7-20 July 1963, Venice.

Torng, H.C.: An approach for the realization of linearily-separable switching functions. IEEE Transactions on Electronic Computers 15:1 (1966 February) 14-20.

U.S. Government Research and Development Reports. Published semi-monthly by the Clearinghouse for Federal Scientific and Technical Information, Springfield, Va. (Title changed 1971 March 25 to: Government Reports Announcements. Now published by the National Technical Information Service.)

U.S. Government Research and Development Reports Annual Subject Index 69:1-24 (1969 January-December). Clearinghouse for Federal Scientific and Technical Information, Springfield, Va. (Title changed 1971 March 25 to: Government Reports Index. Now published by the National Technical Information Service.)

Wilson, Patrick: Two kinds of power, an essay on bibliographical control. University of California Press, Berkeley (1968) 69-92.

Wyllys, R.E.: Research in techniques for improving automatic abstracting procedures. System Development Corp., Santa Monica, Calif. (1963 April 19) Report TM-1087/000/01.

Zunde, P.: Automatic indexing from machine readable abstracts of scientific documents. Documentation Inc., Bethesda, Md. (1965 September) AD 481 148.